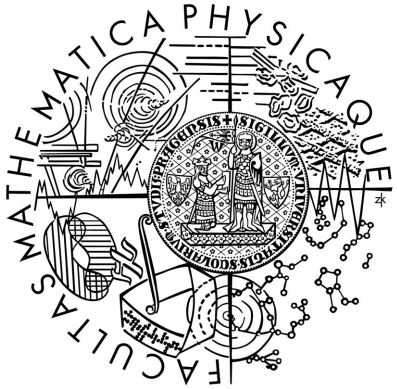
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Bc. Štěpán Škoda

Modelování četností pojistných událostí

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Martin Branda, Ph.D

Studijní program: Matematika

Studijní obor: Finanční a pojistná matematika

Praha 2013

Rád bych poděkoval svému vedoucímu práce Martinu Brandovi za trpělivost, vstřícné jednání a poskytnutí materiálů, dále Martinu Vachatovi za korekturu, společnosti Kooperativa za poskytnutí dat a především mé rodině za finanční podporu při studiích.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 16.4 2013 Štěpán Škoda

Název práce: Modelování četností pojistných událostí

Autor: Bc. Štěpán Škoda

Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Martin Branda, Ph.D, katedra pravděpodobnosti a matematické statistiky

Abstrakt: Předložená práce se zabývá studiem technik pro odhad rizikovosti klientů z hlediska počtů způsobených pojistných událostí, a to na základě údajů, které jsou obsaženy v pojistných smlouvách. Na počátku se blíže věnuje teorii zobecněných lineárních modelů, které mají v oblasti pojišťovnictví široké spektrum využití. V druhé kapitole je představen základní model poissonovské regrese a dále také některé verifikační metody. Speciálně je zde uveden devianční a Waldův test, ale také výsledky platné pro rezidua. Třetí kapitola obsahuje informace o alternativních přístupech k modelování četností pojistných událostí a v závěru je popsána metoda GEE, kterou lze aplikovat, má-li uživatel k dispozici panelová data. Nakonec jsou v numerické studii na konkrétním příkladě ilustrovány popsané techniky, přičemž jako nástroj byl využit statistický software SAS.

Klíčová slova: Zobecněný lineární model, poissonovská regrese, nadprůměrná disperze, GEE metoda

Title: Claims count modeling in insurance

Author: Bc. Štěpán Škoda

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Martin Branda, Ph.D, department of Probability and Mathematical Statistics

Abstract: The present work investigates techniques of insurence ratemaking according to the claims counts of policyholders on the basis of information contained in policies. At the beginning, we provide a closer examination of the theory of generalized linear models, which have wide range of applications in the field of actuarial modeling. The second chapter presents the basic Poisson regression model as well as some particular verification methods. Specifically, deviance and Wald test could be found here and furthermore also important results for residuals. The third chapter contains information on alternative approaches to modeling the claim frequencies and at the end the GEE method, that can be applied in case of panel data, is described. The numerical study based on real insurace data in last part of this diploma thesis illustrate's previously described techniques which were obtained with the help of statistical software SAS.

Keywords: Generalized linear models, Poisson regression, overdispersion, GEE method

Obsah

Úvod

Moderní techniky v oblasti pojišťovnictví využívají pro odhad rizikovosti klienta údajů obsažených v pojistné smlouvě. Příhodný matematický nástroj pro tyto účely tvoří tzv. zobecněné lineární modely. Ty umožňují identifikovat, jaký mají vybrané ukazatele vliv na rizikovost, která je zde vyjádřena odhadem střední hodnoty počtu pojistných událostí v určitém období. Získané výsledky pak mohou dále hrát důležitou roli při stanovení výše pojistného.

V první kapitole se budeme zabývat obecnou teorií, která představuje nadstavbu klasické lineární regrese, a jejíž znalost je pro porozumění textu fundamentální. Navíc zdejší výklad předpokládá, že čtenář absolvoval alespoň nějaký ze základních kurzů statistiky. V příloze A jsou pro úplnost uvedeny poznatky týkající se standardní teorie maximální věrohodnosti, které by též měly být apriorně známy.

Následující dvě kapitoly se blíže věnují modelování četností pojistných událostí. Především je zde představen základní přístup založený na poissonovské regresi. Dále jsou také uvedeny i jeho možné alternativy, které lze využít, pokud není splněn elementární požadavek rovnosti střední hodnoty a rozptylu. Závěrečná sekce třetí kapitoly obsahuje informace o metodě GEE, která umožňuje uživateli vycházet z panelových dat, neboť nevyžaduje předpoklad nezávislosti.

Nakonec poslední čtvrtá kapitola nazvaná numerická studie ilustruje techniky pro identifikaci vhodného modelu na základě výstupů získaných pomocí statistického softwaru SAS, přičemž je vycházeno z reálných dat poskytnutých pojišťovnou Kooperativa. Na závěr je pak porovnáno několik modelů z hlediska jejich predikční schopnosti.

Z dostupné literatury pravděpodobně nejlépe pokrývá v této práci uvedené poznatky kniha [5], kterou bych čtenáři se zájmem o danou problematiku doporučil na prvním místě. První kapitola byla sestavena zejména na základě [9] a [8] a ve zbytku jsem primárně čerpal z [4].

Kapitola 1

Základy teorie zobecněných lineárních modelů

Z pohledu historie matematiky je teorie zobecněných lineárních modelů poměrně novodobou záležitostí. Poprvé byla prezentována v článku *Generalized Linear Models* z roku 1972 a ukázala, jak sjednotit a navíc také dále přirozeně rozšířit do té doby známé přístupy k vyváření regresních modelů pomocí lineární, logistické nebo poissonovské regrese. V současnosti nachází mnohá uplatnění zejména v oblasti pojišťovnictví, biologie, ekonomie a medicíny, ale můžeme se setkat i s aplikacemi v dalších vědních oborech.

Téměř veškeré odhadové a verifikační techniky jsou zde založeny na závěrech teorie maximální věrohodnosti. Proto jsou v příloze zopakovány základní poznatky, se kterými by měl být čtenář obeznámen. V případě potřeby hlubšího porozumění doporučuji čtenáři nahlédnout nejprve do [1], kde je daná problematika popsána podrobněji včetně patřičných důkazů, a odkud také bylo čerpáno. Obsáhleji než [1] se poté zabývá teorií odhadu kniha [12], v níž je možno dohledat doplňující informace.

1.1 Hustota exponenciálního typu

Jedním ze základních pojmů teorie zobecněných lineárních modelů je rozdělení exponenciálního typu, jež nahrazuje funkci normálního rozdělení v lineární regresi. Jde o velmi obecný pojem, díky kterému lze modelovat chování střední hodnoty jak spojitých tak diskrétních náhodných veličin či obecněji vektorů. Následující kapitola se zabývá pouze jednorozměrným případem, zobecnění je popsáno v dizertační práci [8]. V této kapitole bylo také čerpáno z [9].

Třída rozdělení exponenciálního typu zahrnuje širokou škálu rozdělení, mezi něž patří například normální, Poissonovo, gama, alternativní, geometrické či inverzní Gaussovo. Známe-li apriorně hodnotu některých parametrů rozdělení, pak do výčtu dle [3] (str. 21) můžeme také přidat Weibullovo, binomické, negativně binomické nebo Paretovo rozdělení, obecně však toto provést nelze.

Definice 1.1. *Řekneme, že reálná náhodná veličina Y má rozdělení exponenciálního typu, existují-li borelovské funkce $b : \Theta \to \mathbb{R}$, $c : \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ takové, že hustotu*

Y vzhledem k nějaké σ-konečné míře μ lze pro nějaké $\theta \in \Theta \subset \mathbb{R}$, $\varphi \in \mathbb{R}^+$ a $\tilde{w} \in \mathbb{R}^+$ zapsat ve tvaru

$$f(y;\theta,\varphi,\tilde{w}) = \exp\left(\tilde{w}\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi,\tilde{w})\right) \tag{1.1.1}$$

a systém

$$\left\{\exp\left(\tilde{w}\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi,\tilde{w})\right) : \theta \in \Theta \subset \mathbb{R}, \varphi \in \mathbb{R}^+, \tilde{w} \in \mathbb{R}^+\right\}$$

obsahuje hustoty reálných náhodných veličin vzhledem k μ. Parametr θ nazýváme kanonickým a φ disperzním. Argument $\tilde{w}$ nazýváme vahou. O hustotách náhodných veličin s rozdělením exponenciálního typu mluvíme jako o hustotách exponenciálního typu a jejich zápisu ve tvaru (1.1.1) říkáme kanonický.

Tvrzení 1.2. Nechť má náhodná veličina Y rozdělení exponenciálního typu s hustotou v kanonickém tvaru a funkce b je dvakrát spojitě diferencovatelná na Θ (tj. $b \in C^2(\Theta)$), kde Θ je neprázdná otevřená podmnožina $\mathbb{R}$. Potom momentová vytvořující funkce $M_Y(t) = \mathrm{E}\, e^{tY}$ je dvakrát diferencovatelná v bodě nula a navíc platí:

$$\mathrm{E}\, Y = b'(\theta), \qquad \mathrm{var}\, Y = \frac{\varphi}{\tilde{w}} b''(\theta). \tag{1.1.2}$$

Důkaz. Z otevřenosti množiny Θ plyne existence $\varepsilon > 0$ takového, že pro okolí $\omega = (\theta - \varepsilon, \theta + \varepsilon)$ bodu θ platí $\omega \subset \Theta$. Pro t, která splňují podmínku $t\frac{\varphi}{\tilde{w}} + \theta \in \omega$, můžeme provést následující úpravy:

$$\begin{aligned}
\mathrm{E}\, e^{tY} &= \int_{-\infty}^{\infty} \exp\left(\frac{y(t\frac{\varphi}{\tilde{w}} + \theta) - b(\theta)}{\varphi/\tilde{w}} + c(y,\varphi,\tilde{w})\right) \mathrm{d}\mu(y) \\
&= \exp\left(\frac{b\left(t\frac{\varphi}{\tilde{w}} + \theta\right) - b(\theta)}{\varphi/\tilde{w}}\right) \cdot \\
&\quad \cdot \underbrace{\int_{-\infty}^{\infty} \exp\left(\frac{y(t\frac{\varphi}{\tilde{w}} + \theta) - b\left(t\frac{\varphi}{\tilde{w}} + \theta\right)}{\varphi/\tilde{w}} + c(y,\varphi,\tilde{w})\right) \mathrm{d}\mu(y)}_{1} \\
&= \exp\left(\frac{b(t\frac{\varphi}{\tilde{w}} + \theta) - b(\theta)}{\varphi/\tilde{w}}\right) = M_Y(t).
\end{aligned}$$

Momentová vytvořující funkce je zřejmě díky předpokladu $b \in C^2(\Theta)$ pro všechna $t \in (-\varepsilon\frac{\tilde{w}}{\varphi}, \varepsilon\frac{\tilde{w}}{\varphi})$ konečná a i dvakrát diferencovatelná. Zderivováním vytvořující funkce dostáváme:

$$\begin{aligned}
M'_Y(t) &= M_Y(t) b'(t\varphi/\tilde{w} + \theta), \\
M''_Y(t) &= M'_Y(t) b'(t\varphi/\tilde{w} + \theta) + M_Y(t) b''(t\varphi/\tilde{w} + \theta).
\end{aligned}$$

Momenty náhodné veličiny Y spočítáme na základě vlastností momentové vytvořující funkce (viz [5] odstavec 2.3.1), tj.

$$\begin{aligned} \mathrm{E}\, Y &= M_Y'(0) = b'(\theta), \\ \mathrm{E}\, Y^2 &= M_Y''(0) = [b'(\theta)]^2 + \frac{b''(\theta)\varphi}{\tilde{w}}, \\ \mathrm{var}\, Y &= \mathrm{E}\, Y^2 - (\mathrm{E}\, Y)^2 = \frac{\varphi}{\tilde{w}} b''(\theta). \end{aligned}$$

□

Tvrzení 1.3. Mějme systém hustot exponenciálních typů

$$\mathcal{F} = \{f(y;\theta) : \theta \in \Theta\} = \left\{ \exp\left(\tilde{w} \frac{y\theta - b(\theta)}{\varphi} + c(y, \varphi, \tilde{w}) \right) : \theta \in \Theta \right\},$$

kde parametry $\varphi, \tilde{w}$ jsou společné pro všechny hustoty systému. Nechť množina $\Theta \subseteq \mathbb{R}$ je neprázdný otevřený interval, $b \in C^2(\Theta)$ a $b''(\theta) > 0$ pro všechna $\theta \in \Theta$. Pak

- $\mathcal{F}$ je regulární systém hustot s Fisherovou informací $\mathcal{I}(\theta) = \frac{\tilde{w}}{\varphi} b''(\theta)$ pro všechna $\theta \in \Theta$ (viz definice (A.1)),
- funkce $b'(\theta)$ na Θ s oborem hodnot $\{b'(\theta) : \theta \in \Theta\}$ je prostá.

Důkaz. Ověříme postupně předpoklady definice regularity (viz body 1-6 v (A.1)). Bod 1 ($\emptyset \neq \Theta \subseteq \mathbb{R}$, Θ je otevřená množina) je splněn triviálně, 2 ($M = \{y : f(y;\theta) > 0\}$ nezávisí na θ) plyne z definice hustoty exponenciálního typu, která je kladná pro všechna $\theta \in \Theta$, tj. $M = \mathbb{R}$. Derivace $f'(y;\theta) = \frac{\partial f(y;\theta)}{\partial \theta}$ existuje pro všechna $y \in \mathbb{R}$ díky diferencovatelnosti $b(\theta)$ a nezávislosti funkce c na θ (tj. platí také 3). Využitím tvrzení (1.2) při integraci $f'(y;\theta)$ dostáváme:

$$\begin{aligned} \int_{-\infty}^{\infty} f'(y;\theta) \mathrm{d}\mu(y) &= \int_{-\infty}^{\infty} f(y;\theta) \frac{\tilde{w}}{\varphi} (y - b'(\theta)) \mathrm{d}\mu(y) \\ = \frac{\tilde{w}}{\varphi} \underbrace{\int_{-\infty}^{\infty} y f(y;\theta) \mathrm{d}\mu(y)}_{b'(\theta)} &- \frac{\tilde{w}}{\varphi} b'(\theta) \underbrace{\int_{-\infty}^{\infty} f(y;\theta) \mathrm{d}\mu(y)}_{1} = 0 \end{aligned}$$

(tj. platí 4, $\int_M f'(y,\theta) \mathrm{d}\mu(y) = 0$ pro všechna $\theta \in \Theta$). Obdobný výsledek platí také pro $f''(y,\theta) = \frac{\partial^2 f(y,\theta)}{\partial^2 \theta}$,

$$\begin{aligned} \int_{-\infty}^{\infty} f''(y;\theta) \mathrm{d}\mu(y) &= \frac{\tilde{w}}{\varphi} \int_{-\infty}^{\infty} \left[f(y;\theta) \frac{\tilde{w}}{\varphi} (y - b'(\theta)) \right] (y - b'(\theta)) - f(y;\theta) b''(\theta) \mathrm{d}\mu(y) \\ &= \frac{\tilde{w}^2}{\varphi^2} \frac{\varphi}{\tilde{w}} b''(\theta) - \frac{\tilde{w}}{\varphi} b''(\theta) = 0. \end{aligned}$$

Díky $f''(y,\theta) = 0$ platí následující rovnost (viz důkaz věty (A.4)) :

$$\underbrace{\int_{-\infty}^{\infty} \frac{f'(y;\theta) f'(y;\theta)}{f^2(y;\theta)} f(y;\theta) \mathrm{d}\mu(y)}_{\mathcal{I}(\theta)} = - \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(y;\theta)}{\partial^2 \theta} f(y;\theta) \mathrm{d}\mu(y).$$

Nakonec úpravou získáváme pro Fisherovu informaci $\mathcal{I}(\theta)$ vyjádření

$$\mathcal{I}(\theta) = -\int_{-\infty}^{\infty} \frac{\partial^2 \ln f(y;\theta)}{\partial^2 \theta} f(y,\theta)\mathrm{d}\mu(y) = -\int_{-\infty}^{\infty} \left(-\frac{\tilde{w}}{\varphi} b''(\theta)\right) f(y;\theta)\mathrm{d}\mu(y)$$
$$= \frac{\tilde{w}}{\varphi} b''(\theta).$$

Tím máme dokázánu regularitu hustoty $f(y;\theta)$ (tj. platí také 5,6 ($\mathcal{I}(\theta)$ existuje a je kladná)), neboť $b''(\theta)$ je kladné pro všechna $\theta \in \Theta$. Fisherova informace je rovna $\frac{\tilde{w}}{\varphi} b''(\theta)$.

Druhá část tvrzení je přímým důsledkem toho, že funkce b'' je kladná na otevřeném intervalu Θ, a proto je b' rostoucí a zároveň tedy prostá na Θ. □

Povšimněme si, jakým způsobem pak závisí rozptyl na střední hodnotě,

$$\mathrm{E}\, Y = \mu = b'(\theta) \Longrightarrow (b')^{-1}(\mu) = \theta,$$

$$\mathrm{var}\, Y = \frac{\varphi}{\tilde{w}} b''(\theta) = \frac{\varphi}{\tilde{w}} V(\mu) \Longrightarrow V(\mu) = b''((b')^{-1}(\mu)). \tag{1.1.3}$$

Až na konstantu $\varphi/\tilde{w}$ je rozptyl určen střední hodnotou a funkcí V, kterou nazýváme rozptylovou funkcí.

1.1.1 Příklady hustot exponenciálních typů

Ukažme jakým způsobem lze v několika případech převést hustotu do kanonického tvaru, spočtěme rozptylovou funkci a ověřme, že rozptyly a střední hodnoty získané na základě (1.1.2) odpovídají nám známým výsledkům. Níže uvedené příklady byly odvozeny na základě poznatků uvedených v sekci 3.4 knihy [15].

1. Normální rozdělení, $Y \sim N(\mu, \sigma^2)$

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \underbrace{\frac{y^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)}_{c(y,\varphi,\tilde{w})}\right)$$
$$= \exp\left(\tilde{w}\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi,\tilde{w})\right)$$

- Volba parametrů: $\theta = \mu$, $\varphi = \sigma^2$, $\tilde{w} = 1$, $b(\theta) = \frac{\theta^2}{2}$
- Střední hodnota: $\mathrm{E}\, Y = b'(\theta) = \theta = \mu$
- Rozptyl: $\mathrm{var}\, Y = \varphi b''(\theta) = \varphi = \sigma^2$
- Rozptylová funkce: $V(\mu) = 1$

2. Poissonovo rozdělení, $Y \sim Po(\lambda)$

$$f(y;\lambda) = \frac{\lambda^y}{y!}e^{-\lambda} = \exp(y\log\lambda - \lambda - \underbrace{\log y!}_{c(y,\varphi,\tilde{w})}), \qquad y \in \mathbb{N}_0, \lambda > 0$$

- Volba parametrů: $\theta = \log\lambda$, $\varphi = 1$, $\tilde{w} = 1$, $b(\theta) = e^\theta$
- Střední hodnota: $\mathrm{E}\, Y = b'(\theta) = e^\theta = \lambda$
- Rozptyl: $\mathrm{var}\, Y = \varphi b''(\theta) = \lambda$
- Rozptylová funkce: $V(\mu) = \mu$

3. Gama rozdělení, $Y \sim \Gamma(a,p)$

$$f(y;a,p) = \frac{a^p}{\Gamma(p)}y^{p-1}e^{-ap}, \qquad y > 0, a, p > 0$$

$$\begin{aligned} f(y;a,p) &= \exp(-ay + (p-1)\log y + p\log a - \log\Gamma(p)) \\ &= \exp\left(\frac{y(-\frac{a}{p}) + \log\frac{a}{p}}{\frac{1}{p}} + \underbrace{p(\log p) + (p-1)\log y - \log\Gamma(p)}_{c(y,\varphi,\tilde{w})}\right) \end{aligned}$$

- Volba parametrů: $\theta = -\frac{a}{p}$, $\varphi = \frac{1}{p}$, $\tilde{w} = 1$, $b(\theta) = -\log(-\theta)$
- Střední hodnota: $\mathrm{E}\, Y = b'(\theta) = -\frac{1}{\theta} = \frac{p}{a}$
- Rozptyl: $\mathrm{var}\, Y = \varphi b''(\theta) = \frac{p}{a^2}$
- Rozptylová funkce: $V(\mu) = \mu^2$

Poznámka. Uvědomme si, že $c(y,\varphi,\tilde{w})$ existuje nejen pro kladná y, nýbrž pro každé $y \in \mathbb{R}$. Je-li λ Lebesguevovou mírou na $(\mathbb{R}, \mathcal{B}_0)$, pak příslušná hustota exponenciálního typu existuje vzhledem k σ-konečné míře splňující $\mu(B) = \lambda(B \cap \mathbb{R}^+)$ pro každé $B \in \mathcal{B}_0$.

4. Alternativní rozdělení, $Y \sim Alt(p)$

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

- Volba parametrů: $\theta = \log\frac{p}{1-p}$, $\varphi = 1$, $\tilde{w} = 1$, $b(\theta) = \log(1+e^\theta)$
- Střední hodnota: $\mathrm{E}\, Y = b'(\theta) = \frac{e^\theta}{1+e^\theta} = p$
- Rozptyl: $\mathrm{var}\, Y = \varphi b''(\theta) = p - p^2 = p(1-p)$
- Rozptylová funkce: $V(\mu) = \mu(1-\mu)$

5. Negativně binomické při známém pevném parametru $a > 0$, $Y \sim NB(a, \lambda)$

$$f(y;\lambda) = \binom{a+y-1}{y}\left(\frac{\lambda}{a+\lambda}\right)^y\left(\frac{a}{a+\lambda}\right)^a, \qquad y \in \mathbb{N}_0, \lambda > 0$$
$$= \exp\left(\log\binom{a+y-1}{y} + y\log\left(\frac{a\lambda}{a+\lambda}\cdot\frac{1}{a}\right) - a\log\left(1+\frac{\lambda}{a}\right)\right)$$
$$= \exp\left(y\log\left(\frac{a\lambda}{a+\lambda}\right) - a\log\left(1+\frac{\lambda}{a}\right) + \underbrace{y\log\left(\frac{1}{a}\right) + \log\binom{a+y-1}{y}}_{c(y,\varphi,\tilde{w})}\right)$$

- Volba parametrů: $\theta = \log(\frac{a\lambda}{a+\lambda})$, $\varphi = 1$, $\tilde{w} = 1$, $b(\theta) = a\log(1+\frac{\exp(\theta)}{a-\exp(\theta)})$
- Střední hodnota: $\mathrm{E}\, Y = b'(\theta) = a\frac{\exp(\theta)}{a-\exp(\theta)} = \lambda$
- Rozptyl: $\mathrm{var}\, Y = \varphi b''(\theta) = \frac{a^2\exp(\theta)}{(a-\exp(\theta))^2} = \lambda\frac{a}{a-\exp(\theta)} = \lambda(1+\frac{\lambda}{a})$
- Rozptylová funkce: $V(\mu) = \mu + \frac{1}{a}\mu^2$

1.2 Formulace modelu

Podobně jako u lineární regrese, také ve zobecněných lineárních modelech (generalized linear models) vystupují nezávislé náhodné veličiny označované zpravidla $Y_1, \ldots, Y_n$ jako vysvětlované proměnné (v literatuře též nazývané odezvy, regresandy či závisle proměnné) s realizací $y_1, \ldots, y_n$ a vektory vysvětlujících proměnných (regresorů, kovariát, prediktorů či nezávisle proměnných) $\mathbf{x}_1, \ldots, \mathbf{x}_n$, o nichž budeme předpokládat, že mají délku p. Účelem modelu je co nejlépe vystihnout vliv vysvětlujících proměnných na proměnnou vysvětlovanou.

Existují dvě možné formulace zobecněného lineárního modelu v závislosti na tom, zda budeme považovat regresory za náhodné vektory nebo vektory konstant. Odhadové a verifikační techniky bývají v obou případech identické a proto volba formulace nemá z praktického hlediska zásadní význam.

1. Regresory jako vektory konstant

Model tvoří tři elementární stavební komponenty:

1. Vysvětlované proměnné $Y_1, \ldots, Y_n$ jsou nezávislé náhodné veličiny s rozdělením exponenciálního typu. Hustota Y_i má tvar

$$f(y_i;\theta_i,\varphi,\tilde{w}_i) = \exp\left(\tilde{w}_i\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i,\varphi,\tilde{w}_i)\right) \text{ pro } i = 1,\ldots,n, \quad (1.2.1)$$

přičemž hodnoty parametrů $\theta_1, \ldots, \theta_n$ pocházejí z otevřeného intervalu $\Theta \subseteq \mathbb{R}$, $b \in C^2(\Theta)$, $b''(\theta) > 0$ pro všechna $\theta \in \Theta$, $\varphi > 0$ je společným disperzním parametrem všech hustot a $\tilde{w}_i > 0$ je apriorně známá váha.

2. Parametr θ_i závisí na $\mathbf{x}_i$ prostřednictvím lineárního prediktoru η_i, pro který platí
$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{1.2.2}$$
kde vektor $\boldsymbol{\beta} \in \mathbf{B} \subseteq \mathbb{R}^p$ obsahuje p neznámých hodnot parametrů a $p \ll n$.

3. Existuje funkce $g : \mu(\Theta) \to \mathbb{R}$, kde $\mu(\Theta) = \{b'(\theta) : \theta \in \Theta\}$, taková, že $g \in C^1(\mu(\Theta))$ a $g'(\mu) \neq 0$, pro všechna $\mu \in \mu(\Theta)$, a pro kterou
$$\eta_i = g(\mu_i),$$
kde $\mu_i = \mathrm{E}\, Y_i$. Tato funkce se nazývá linková.

Tato formulace se z důvodu jednoduchosti zápisu, interpretace a pozdější práce s modelem v literatuře vyskytuje častěji a i my se jí v této práci budeme držet, nebude-li výslovně řečeno jinak. Náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ řídící se zobecněným lineárním modelem s linkovou funkcí g a maticí plánu (design matrix) $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ budeme stručně zapisovat symbolem $\mathbf{Y} \sim GLM(f, g, \mathbf{X})$. Poznamenejme, že pro použití maximálně věrohodných odhadů je nutné teorii uvedenou v příloze A dále rozšířit, aby platila také pro nestejně rozdělené náhodné veličiny (viz věta 1.6).

2. Regresory jako náhodné vektory

Označme p-rozměrné náhodné vektory regresorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Tato formulace se oproti předešlé liší především v prvním bodě, předpokládáme totiž, že celé vektory $(Y_1, \mathbf{X}_1^T)^T, \ldots, (Y_n, \mathbf{X}_n^T)^T$ jsou nezávislé a navíc stejně rozdělené. Hustotu (1.2.1) mají tentokráte náhodné veličiny $(Y_i|\mathbf{X}_i = \mathbf{x}_i)$ a μ_i z bodu 3 poté chápeme jako podmíněnou střední hodnotu $\mathrm{E}\,(Y_i|\mathbf{X}_i = \mathbf{x}_i)$, zbytek se nijak neliší. Obdobně i ostatní charakteristiky závisle proměnné chápeme vždy podmíněně. V tomto případě již lze odhadnout model pomocí „standardní“ teorie maximální věrohodnosti.[1]

Ukažme nyní proč jsou odhady dle obou formulací shodné. Předpokládejme, že φ a $\tilde{w}_1, \ldots, \tilde{w}_n$ jsou apriorně známé hodnoty, přičemž v praxi je často nejprve třeba φ vhodným způsobem odhadnout. Hustotu $f_{(Y_i, \mathbf{X}_i^T)^T}(y_i, \mathbf{x}_i; \theta_i, \varphi, \tilde{w}_i)$ náhodného vektoru $(Y_i, \mathbf{X}_i^T)^T$ můžeme zapsat jako součin podmíněné hustoty $f(y_i|\mathbf{x}_i; \theta_i, \varphi, \tilde{w}_i)$ odpovídající (1.2.1) a hustoty náhodného vektoru regresorů $f_{\mathbf{X}_i}(\mathbf{x}_i)$. Věrohodnostní funkce ℓ má poté tvar
$$\ell(\boldsymbol{\beta}) = \Pi_{i=1}^n f_{(Y_i, \mathbf{X}_i^T)^T}(y_i, \mathbf{x}_i; \theta_i, \varphi, \tilde{w}_i) = \Pi_{i=1}^n f(y_i|\mathbf{x}_i; \theta_i, \varphi, \tilde{w}_i) f_{\mathbf{X}_i}(\mathbf{x}_i)$$
a logaritmická věrohodnostní funkce
$$\ln \ell(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) = \sum_{i=1}^n \log f(y_i|\mathbf{x}_i; \theta_i, \varphi, \tilde{w}_i) + \sum_{i=1}^n \log f_{\mathbf{X}_i}(\mathbf{x}_i).$$

[1] Pro korektní použití teorie je nutné definice a věty uvedené v příloze A modifikovat pro posloupnosti náhodných vektorů.

Zderivováním dle $\boldsymbol{\beta}$ dostáváme

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i|\mathbf{x}_i; \theta_i, \varphi, \tilde{w}_i).$$

To znamená, že věrohodnostní rovnice jsou pro oba typy formulací identické, a proto také odhady parametru $\boldsymbol{\beta}$.

Zobecněný lineární model lze ekvivalentně popsat pomocí čtyřech možných typů parametrizace:

$$\begin{aligned} \boldsymbol{\beta} &= (\beta_1, \ldots, \beta_p)^T, \\ \boldsymbol{\theta} &= (\theta_1, \ldots, \theta_n)^T, \\ \boldsymbol{\eta} &= (\eta_1, \ldots, \eta_n)^T, \\ \boldsymbol{\mu} &= (\mu_1, \ldots, \mu_n)^T. \end{aligned}$$

Vztahy mezi nimi jsou dány definicí modelu a názorně je popisuje následující schéma:

$$\boldsymbol{\beta} \underbrace{\overset{\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}}{\rightarrow}}_{} \boldsymbol{\eta} \underbrace{\underbrace{\overset{\mu_i = g^{-1}(\eta_i)}{\rightleftarrows}}_{\eta_i = g(\mu_i)}}_{} \boldsymbol{\mu} \underbrace{\underbrace{\overset{\theta_i = (b')^{-1}(\mu_i)}{\rightleftarrows}}_{\mu_i = b'(\theta_i)}}_{} \boldsymbol{\theta}$$

Parametrizaci $\boldsymbol{\beta}$ říkáme hlavní. Hustotu vysvětlované proměnné Y_i budeme vzhledem k různým parametrizacím stručně zapisovat ve tvarech: $f(y_i; \boldsymbol{\beta})$, $f(y_i; \theta_i)$, $f(y_i; \eta_i)$ či $f(y_i; \mu_i)$.

1.3 Odhad parametru $\boldsymbol{\beta}$

Vyjádřeme nejprve pomocí řetízkového pravidla tvar skórového vektoru $U(\boldsymbol{\beta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\beta})$, kde

$$U_i(\boldsymbol{\beta}) = \frac{\partial \ln f(Y_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \ln f(Y_i; \theta_i)}{\partial \theta_i} \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{1.3.1}$$

s $\theta_i(\mu_i) = (b')^{-1}(\mu_i)$, $\mu_i(\eta_i) = g^{-1}(\eta_i)$, $\eta_i(\boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta}$, tvoří příspěvek i-tého pozorování k celkové hodnotě $U(\boldsymbol{\beta})$. Úpravami dostáváme

$$\begin{aligned} \frac{\partial \ln f(Y_i; \theta_i)}{\partial \theta_i} &= \frac{\tilde{w}_i}{\varphi}(Y_i - b'(\theta_i)) = \frac{\tilde{w}_i}{\varphi}(Y_i - \mu_i), \\ \frac{\partial \mu_i(\theta_i)}{\partial \theta_i} &= b''(\theta_i) = V(\mu_i) \Rightarrow \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} = \frac{1}{V(\mu_i)}, \qquad (1.3.2) \\ \frac{\partial \eta_i(\mu_i)}{\partial \mu_i} &= g'(\mu_i) \Rightarrow \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} = \frac{1}{g'(\mu_i)}, \qquad (1.3.3) \\ \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \begin{pmatrix} \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_p} \end{pmatrix} = \mathbf{x}_i, \end{aligned}$$

přičemž ve (1.3.2) a (1.3.3) byla užita věta o derivaci inverzní funkce. Existenci parciální derivace

$$\frac{\partial\theta_i(\mu_i)}{\partial\mu_i} = \frac{1}{V(\mu_i)} = \frac{1}{b''((b')^{-1}(\mu_i))}$$

zaručuje předpoklad $b'' > 0$, ze kterého také vyplývá, že b' je rostoucí funkce na Θ. Spojitosti b' na intervalu Θ implikuje, že množina $\mu(\Theta) = \{b'(\theta) : \theta \in \Theta\}$ je interval (Darbouxova vlastnost), a proto parciální derivace

$$\frac{\partial\mu_i(\eta_i)}{\partial\eta_i} = \frac{1}{g'(\mu_i)}$$

existuje díky předpokladům $g \in C^1(\mu(\Theta))$, $g'(\mu) \neq 0$ pro $\mu \in \mu(\Theta)$, z kterých plyne ryzí monotonie funkce g na $\mu(\Theta)$. Vyjádření má tedy tvar

$$U_i(\boldsymbol{\beta}) = \frac{\tilde{w}_i}{\varphi}(Y_i - \mu_i)\frac{\mathbf{x}_i}{V(\mu_i)g'(\mu_i)}.$$

Označme

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

předpis funkce w, kterou zřejmě můžeme takto definovat na celém $\mu(\Theta)$. Skórový vektor je pak roven

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\beta}) = \frac{1}{\varphi}\sum_{i=1}^{n} \tilde{w}_i w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\mathbf{x}_i. \tag{1.3.4}$$

Poznamenejme, že z předpokladu $\mu_i = \mathrm{E}\, Y_i$ zřejmě plyne

$$\mathrm{E}\, U(\boldsymbol{\beta}) = \mathrm{E}\sum_{i=1}^{n} U_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{E}\, U_i(\boldsymbol{\beta}) = 0.$$

Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ je kořenem soustavy rovnic $U(\boldsymbol{\beta}) = 0$. Tj., pro $\hat{\boldsymbol{\beta}}$ je

$$\sum_{i=1}^{n} \tilde{w}_i w(\hat{\mu}_i) g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)\mathbf{x}_i = 0,$$

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

Definice 1.4. *Linkovou funkci g nazýváme kanonická právě tehdy, když je inverzní funkcí b', tj. $g = (b')^{-1}$.*

Existence kanonické linkové funkce je zaručena díky tvrzení 1.3. Jestliže je g kanonická linková funkce, pak platí:

$$\begin{aligned} b'(\theta_i) = \mu_i \;\wedge\; g(\mu_i) = (b')^{-1}(\mu_i) &\Rightarrow \theta_i = g(\mu_i), \\ (b')^{-1}(\mu_i) = \theta_i \;\wedge\; b''(\theta_i) = V(\mu_i) &\Rightarrow g'(\mu_i) = \frac{1}{V(\mu_i)}. \end{aligned} \tag{1.3.5}$$

Implikace na druhém řádku plyne z věty o derivaci inverzní funkce. Využitím těchto vztahů dojde ke značnému zjednodušení vyjádření skórového vektoru, neboť přejde v

$$U(\boldsymbol{\beta}) = \frac{1}{\varphi}\sum_{i=1}^{n} w_i(Y_i - \mu_i)\mathbf{x}_i.$$

Parametr $\boldsymbol{\beta}$ se pak odhadne řešením rovnice $\sum_{i=1}^{n} \tilde{w}_i(Y_i - \hat{\mu}_i)\mathbf{x}_i = 0$, tj., pro $\hat{\boldsymbol{\beta}}$ platí

$$\sum_{i=1}^{n} \tilde{w}_i Y_i \mathbf{x}_i = \sum_{i=1}^{n} \tilde{w}_i \hat{\mu}_i \mathbf{x}_i = \sum_{i=1}^{n} \tilde{w}_i g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})\mathbf{x}_i. \tag{1.3.6}$$

Ekvivalentně lze také užít maticový zápis

$$\mathbf{X}^T\tilde{\mathbf{W}}\mathbf{Y} = \mathbf{X}^T\tilde{\mathbf{W}}\hat{\boldsymbol{\mu}},$$

kde $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ je matice typu $n \times p$, $\tilde{\mathbf{W}} = diag(\tilde{w}_1, \ldots, \tilde{w}_n)$ diagonální matice vah, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ a $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^T$.

1.4 Numerické řešení věrohodnostních rovnic

Obecně není pro zobecněné lineární modely k dispozici jednoduché explicitní řešení věrohodnostních rovnic, a proto software užívá numerické algoritmy. Standardně se v literatuře můžeme setkat s metodou iterativních vážených nejmenších čtverců (IWLS, iterated weighted least squares) a Newtonovým–Raphsonovým algoritmem. Ve skutečnosti jsou oba tyto přístupy založeny na stejném principu a navíc při volbě odpovídajících počátečních hodnot a kanonické linkové funkce jejich iterace splývají.

Newton–Raphsonův algoritmus

Nejprve se pro $k = 0$ zvolí vhodná počáteční hodnota $\hat{\boldsymbol{\beta}}^{(0)}$ a poté opakujeme následující dva kroky, dokud není splněno kritérium konvergence $\| \hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)} \| < \varepsilon$.

1. Spočteme odhad skórové funkce $U(\hat{\boldsymbol{\beta}}^{(k)})$ (viz (1.3.4)) a výběrovou informační matici $F_n(\hat{\boldsymbol{\beta}}^{(k)}) = nF(\hat{\boldsymbol{\beta}}^{(k)})$ (viz (A.3.1)) na základě $\hat{\boldsymbol{\beta}}^{(k)}$.

2. Nový odhad parametru určíme dle vyjádření

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - [F_n(\hat{\boldsymbol{\beta}}^{(k)})]^{-1} U(\hat{\boldsymbol{\beta}}^{(k)})$$

   a hodnotu k navýšíme o 1.

Jako počáteční hodnota $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_1^{(0)}, \ldots, \hat{\beta}_n^{(0)})^T$ se například někdy užívá vektor se složkami

$$\hat{\beta}_1^{(0)} = g\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \text{ a } \hat{\beta}_j^{(0)} = 0, j = 2, \ldots, p.$$

Nyní ukážeme, proč mají kanonické linkové funkce výsadní postavení v rámci zobecněných lineárních modelů, přičemž využijeme vyjádření výběrové informační matice. Poté ještě před popisem samotného průběhu metody iterativních vážených nejmenších čtverců plynně navážeme odvozením jejího principu, z čehož bude nakonec patrná souvislost s Newton-Raphsonovým algoritmem.

Existuje-li Fisherova informační matice $\mathcal{I}_n(\boldsymbol{\beta})$, pak lze zapsat ve tvaru (viz (A.1.6)):

$$\begin{aligned}\mathcal{I}_n(\boldsymbol{\beta}) = \operatorname{var} U(\boldsymbol{\beta}) &= \operatorname{var} \frac{1}{\varphi}\sum_{i=1}^{n} \tilde{w}_i w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\mathbf{x}_i \\ &= \frac{1}{\varphi^2}\sum_{i=1}^{n} \tilde{w}_i^2 w^2(\mu_i)[g'(\mu_i)]^2 \mathbf{x}_i \operatorname{var}(Y_i - \mu_i)\mathbf{x}_i^T \\ &= \frac{1}{\varphi^2}\sum_{i=1}^{n} \frac{\tilde{w}_i^2}{V^2(\mu_i)[g'(\mu_i)]^4}[g'(\mu_i)]^2 \mathbf{x}_i \frac{\varphi}{\tilde{w}_i} V(\mu_i)\mathbf{x}_i^T \\ &= \frac{1}{\varphi}\sum_{i=1}^{n} \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T. \end{aligned} \tag{1.4.1}$$

Při úpravách bylo využito předpokladu nezávislosti mezi vysvětlovanými proměnnými.

Předpokládejme nyní, že g' a w jsou diferencovatelné na $\mu(\Theta)$. Z definice výběrové informační matice $F(\boldsymbol{\beta})$ (viz (A.3.1)), řetízkového pravidla a vztahů pro $U(\boldsymbol{\beta})$ v sekci (1.3) dostáváme:

$$\begin{aligned} F(\boldsymbol{\beta}) &= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ln f(Y_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial U_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T} = -\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial U_i(\boldsymbol{\beta})}{\partial\beta_1}, \ldots, \frac{\partial U_i(\boldsymbol{\beta})}{\partial\beta_p}\right) \\ &= -\frac{1}{\varphi n}\sum_{i=1}^{n} \frac{\tilde{w}_i w(\mu_i) g'(\mu_i)(Y_i - \mu_i)}{\partial\mu_i} \frac{\partial\mu_i(\eta_i)}{\partial\eta_i}\mathbf{x}_i \frac{\partial\eta_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T} \\ = -\frac{1}{\varphi n}\sum_{i=1}^{n} \tilde{w}_i [w'(\mu_i)g'(\mu_i)(Y_i - \mu_i) &+ w(\mu_i)g''(\mu_i)(Y_i - \mu_i) - w(\mu_i)g'(\mu_i)] \frac{1}{g'(\mu_i)}\mathbf{x}_i\mathbf{x}_i^T \\ = \frac{1}{\varphi n}\sum_{i=1}^{n} \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T &- \underbrace{\frac{1}{\varphi n}\sum_{i=1}^{n} \tilde{w}_i \left[w'(\mu_i) + w(\mu_i)\frac{g''(\mu_i)}{g'(\mu_i)}\right](Y_i - \mu_i)\mathbf{x}_i\mathbf{x}_i^T}_{J_n(\boldsymbol{\beta})}. \end{aligned} \tag{1.4.2}$$

Jelikož je $\mathrm{E}\, Y_i = \mu_i$ a tedy také $\mathrm{E}\, J_n(\boldsymbol{\beta}) = 0$, má střední hodnota výběrové informační matice tvar:

$$\mathrm{E}\, F(\boldsymbol{\beta}) = \frac{1}{\varphi n}\sum_{i=1}^{n} \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T.$$

Využitím (1.4.1) za předpokladu existence $\mathcal{I}_n(\boldsymbol{\beta})$ dostáváme:

$$\mathcal{I}_n(\boldsymbol{\beta}) = \mathrm{E}\, nF(\boldsymbol{\beta}) = \frac{1}{\varphi}\sum_{i=1}^{n} \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T.$$

V praxi představuje $nF(\hat{\boldsymbol{\beta}})$ odhad informační matice $\mathcal{I}_n(\boldsymbol{\beta})$, kde $\hat{\boldsymbol{\beta}}$ je kořenem věrohodnostních rovnic, ale na rozdíl od $\mathcal{I}_n(\boldsymbol{\beta})$ nemusí nutně splňovat podmínku pozitivní definitnosti.

Povšimněme si, že předpoklady zobecněného lineárního modelu vyžadují, aby funkce V a $[g']^2$ byly kladné, a proto je také

$$w(\mu_i) = \frac{1}{V(\mu_i)[g'(\mu_i)]^2} > 0 \text{ pro } i = 1, \ldots, n.$$

Z toho dále vyplývá pozitivní semidefinitnost matice $\frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T$, neboť pro libovolný p-rozměrný vektor $\mathbf{c} \in \mathbb{R}^p$ máme

$$\mathbf{c}^T\left(\frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{c} = \frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)\mathbf{c}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{c} = \frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)(\mathbf{x}_i^T\mathbf{c})^2 \geq 0.$$

Navíc platí také následující ekvivalentní výroky:

$$\left(\frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)(\mathbf{x}_i^T\mathbf{c})^2 = 0\right) \Leftrightarrow \left(\sum_{i=1}^n (\mathbf{x}_i^T\mathbf{c})^2 = 0\right) \Leftrightarrow \left(\mathbf{x}_i^T\mathbf{c} = 0 \text{ pro } i = 1, \ldots, n\right) \Leftrightarrow \mathbf{X}\mathbf{c} = 0.$$

To znamená, že matice $\frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T$ je pozitivně definitní právě tehdy, když sloupce $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ jsou lineárně nezávislé.

Je-li $\mathbf{Y} \sim GLM(f, g, \mathbf{X})$ a $\mathbf{x}_i^T\boldsymbol{\beta} \in g(\mu(\Theta)) = \{g(b'(\theta)) : \theta \in \Theta\}$ pro všechna $i = 1 \ldots, n$ a $\boldsymbol{\beta} \in \mathbf{B}$, pak již lze relativně snadno na základě výše zmíněných úvah ukázat, že lineární nezávislost sloupců matice $\mathbf{X}$ je postačující podmínkou pro to, aby hustota náhodného vektoru $\mathbf{Y}$ pocházela z regulárního systému hustot $\{\prod_{i=1}^n f(y_i; \boldsymbol{\beta}); \boldsymbol{\beta} \in \mathbf{B}\}$ s Fisherovou informační maticí $\mathcal{I}_n(\boldsymbol{\beta}) = \frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T$ (viz důsledek 3.9 práce [8]). Tato podmínka ovšem nezaručuje existenci výběrové Fisherovy informační matice, pro jejíž vyjádření jsme potřebovali, aby funkce g' a w byly diferencovatelné.

Předpokládejme nyní, že g je kanonická linková funkce. Z definice 1.4 a vztahu (1.3.5) dostáváme:

$$g(\mu_i) = (b')^{-1}(\mu_i), g'(\mu_i) = \frac{1}{V(\mu_i)} \Rightarrow g''(\mu_i) = -\frac{V'(\mu_i)}{V^2(\mu_i)} \Rightarrow \frac{g''(\mu_i)}{g'(\mu_i)} = -\frac{V'(\mu_i)}{V(\mu_i)}$$
$$w(\mu_i) = \frac{1}{V(\mu_i)[g'(\mu_i)]^2} = V(\mu_i) \Rightarrow w'(\mu_i) = V'(\mu_i). \tag{1.4.3}$$

Část výrazu $J_n(\boldsymbol{\beta})$ (viz (1.4.2)) uvedená v hranaté závorce poté přejde v

$$w'(\mu_i) + w(\mu_i)\frac{g''(\mu_i)}{g'(\mu_i)} = V'(\mu_i) + V(\mu_i)\left(-\frac{V'(\mu_i)}{V(\mu_i)}\right) = 0,$$

a proto se také $J_n(\boldsymbol{\beta})$ rovná 0 a $F(\boldsymbol{\beta}) = \frac{1}{\varphi n}\sum_{i=1}^n \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T$.

Na základě výše uvedených poznatků jsem sestavil následující větu.

Věta 1.5. *Nechť* $\mathbf{Y} \sim GLM(f, g, \mathbf{X})$, $\mathbf{X}$ *má lineárně nezávislé sloupce,* b *je třikrát diferencovatelná na* Θ *a* g *je kanonická linková funkce. Nechť navíc* $\boldsymbol{\beta}_0$ *je skutečná hodnota parametru* $\boldsymbol{\beta}$ *a pochází z otevřeného konvexního parametrického prostoru* $\mathbf{B} \subseteq \mathbb{R}^p$*, přičemž pro všechna* $i = 1, \ldots, n$ *a* $\boldsymbol{\beta} \in \mathbf{B}$ *platí* $\mathbf{x}_i^T\boldsymbol{\beta} \in g(\mu(\Theta)) = \{g(b'(\theta)) : \theta \in \Theta\}$*. Pak, existuje-li pro realizaci* $\mathbf{Y}$ *řešení soustavy věrohodnostních rovnic* $U(\boldsymbol{\beta}) = 0$ *, je určeno jednoznačně a je maximálně věrohodným odhadem parametru* $\boldsymbol{\beta}_0$*.*

Důkaz. Jelikož g je kanonická linková funkce, zaručuje předpoklad existence třetích derivací b diferencovatelnost funkcí g' a w. Z (1.4.3) a (1.1.3) plyne

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

pro všechna $\mu \in \mu(\Theta)$, neboť b'' je kladnou funkcí na Θ. Obdobně můžeme vyjádřit také derivace funkce g ve všech bodech $\mu(\Theta)$. Úpravami dostáváme

$$\begin{aligned} g'(\mu) &= [(b')^{-1}(\mu)]' = \frac{1}{b''\left((b')^{-1}(\mu)\right)}, \\ g''(\mu) &= -\left(\frac{1}{b''\left((b')^{-1}(\mu)\right)}\right)^2 b'''\left((b')^{-1}(\mu)\right)\frac{1}{b''\left((b')^{-1}(\mu)\right)} = -\frac{b'''\left((b')^{-1}(\mu)\right)}{\left(b''\left((b')^{-1}(\mu)\right)\right)^3}. \end{aligned}$$

Z diferencovatelnosti g' a w plyne existence výběrové Fisherovy informační matice (viz (1.4.2)). Díky kanonické linkové funkci, předpokladu $\mathbf{x}_i^T\boldsymbol{\beta} \in g(\mu(\Theta))$ a nezávislosti sloupců $\mathbf{X}$ je

$$\begin{aligned} F_n(\boldsymbol{\beta}) = -\sum_{i=1}^n \frac{\partial^2 \ln f(Y_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w\left(g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})\right)\mathbf{x}_i\mathbf{x}_i^T \\ &= \frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w\left(b'(\mathbf{x}_i^T\boldsymbol{\beta})\right)\mathbf{x}_i\mathbf{x}_i^T \end{aligned}$$

pro všechna $\boldsymbol{\beta} \in \mathbf{B}$ rovna pozitivně definitní Fisherově informační matici $\mathcal{I}_n(\boldsymbol{\beta})$. Spojitost w a b' implikuje spojitost funkce $\boldsymbol{\beta} \mapsto w\left(b'(\mathbf{x}_i^T\boldsymbol{\beta})\right)$ na $\mathbf{B}$. To znamená, že druhé parciální derivace

$$\frac{\partial^2 L}{\partial\beta_r\partial\beta_s}$$

logaritmické věrohodnostní funkce L jsou spojité na $\mathbf{B}$ pro všechna $r, s = 1, \ldots, p$, neboť platí

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_r\partial\beta_s} = \frac{\partial^2\sum_{i=1}^n \ln f(Y_i;\boldsymbol{\beta})}{\partial\beta_r\partial\beta_s} = \sum_{i=1}^n \frac{\partial^2 \ln f(Y_i;\boldsymbol{\beta})}{\partial\beta_r\partial\beta_s} = -\frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w\left(b'(\mathbf{x}_i^T\boldsymbol{\beta})\right)x_{ir}x_{is}.$$

Dle věty 1.74 textu [11] je pak logaritmická věrohodnostní funkce ryze konkávní na konvexní množině, a proto se v každém řešení soustavy věrohodnostních rovnic nachází lokální maximum, které je jednoznačně určeným globálním maximem (viz [11] věta 1.66)). □

Vraťme se zpět k metodě iterativních vážených nejmenších čtverců. Abychom ukázali souvislost s Newtonovým–Raphsonovým algoritmem, vyjděme z rovnice

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - [F_n(\hat{\boldsymbol{\beta}})]^{-1} U(\hat{\boldsymbol{\beta}}),$$

kde $\hat{\boldsymbol{\beta}}$ je maximálně věrohodný odhad $\boldsymbol{\beta}$ pro realizaci $\mathbf{Y}$, tj. $U(\hat{\boldsymbol{\beta}}) = 0$, a $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$. Výběrovou Fisherovu matici $F_n(\hat{\boldsymbol{\beta}})$ nahradíme

$$\mathcal{I}_n(\hat{\boldsymbol{\beta}}) = \frac{1}{\varphi} \sum_{i=1}^{n} \tilde{w}_i w(\hat{\mu}_i) \mathbf{x}_i \mathbf{x}_i^T \tag{1.4.4}$$

a vynásobíme jí zleva obě strany rovnice. Tím dostáváme

$$\mathcal{I}_n(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\beta}} = \mathcal{I}_n(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{\beta}} + U(\hat{\boldsymbol{\beta}}),$$

a tedy

$$\begin{aligned}
\frac{1}{\varphi} \sum_{i=1}^{n} \tilde{w}_i w(\hat{\mu}_i) \mathbf{x}_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}} &= \frac{1}{\varphi} \sum_{i=1}^{n} \tilde{w}_i w(\hat{\mu}_i) \mathbf{x}_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \frac{1}{\varphi} \sum_{i=1}^{n} \tilde{w}_i w(\hat{\mu}_i) g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i) \mathbf{x}_i \\
(\sum_{i=1}^{n} \tilde{w}_i w(\hat{\mu}_i) \mathbf{x}_i \mathbf{x}_i^T) \hat{\boldsymbol{\beta}} &= \sum_{i=1}^{n} \tilde{w}_i w(\hat{\mu}_i) \mathbf{x}_i \underbrace{[\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)]}_{\hat{Z}_i}.
\end{aligned}$$

Poslední rovnici můžeme zapsat přehledněji pomocí maticového zápisu

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

kde $\hat{\mathbf{W}} = diag(\tilde{w}_1 w(\hat{\mu}_1), \ldots, \tilde{w}_n w(\hat{\mu}_n))$ je diagonální matice $n \times n$ a $\hat{\mathbf{Z}} = (\hat{Z}_1, \ldots, \hat{Z}_n)^T$ n-rozměrný vektor, který bývá nazýván linearizovaná odezva. Zpětným vyjádřením $\hat{\boldsymbol{\beta}}$ dostáváme:

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

Metoda iterativních vážených nejmenších čtverců (IWLS)

1. Vezmeme počáteční odhad $\hat{\boldsymbol{\mu}}^{(0)} = \mathbf{Y}$ a hodnotu k nastavíme na 0.

2. Spočteme

$$\begin{aligned}
\hat{\mathbf{W}}^{(k)} &= diag(\tilde{w}_1 w(\hat{\mu}_1^{(k)}), \ldots, \tilde{w}_n w(\hat{\mu}_n^{(k)})), \\
\hat{Z}_i^{(k)} &= g(\hat{\mu}_i^{(k)}) + g'(\hat{\mu}_i^{(k)})(Y_i - \hat{\mu}_i^{(k)}) \text{ pro } i = 1, \ldots n,
\end{aligned}$$

   a přejdeme k bodu 3.

3. Započneme iteraci vyjádřením

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}^T \hat{\mathbf{W}}^{(k)} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}}^{(k)} \hat{\mathbf{Z}}^{(k)}, \tag{1.4.5}$$

   kde $\hat{\mathbf{Z}}^{(k)} = (\hat{Z}_1^{(k)} \ldots, \hat{Z}_n^{(k)})^T$. Pokud je splněno kritérium konvergence $\| \hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)} \| < \varepsilon$, pak algoritmus končí, jinak pokračujeme bodem 4.

4. Dopočteme $\mu_i^{(k+1)} = g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(k+1)})$ pro $i = 1, \dots n$, navýšíme hodnotu k o 1 a přejdeme k bodu 2.

Z výše uvedených poznatků plyne, že pro kanonickou linkovou funkci jsou obě výše uvedené numerické metody (IWLS, Newton–Raphson) ekvivalentní ve smyslu identičnosti iterací, a to v důsledku rovnosti výběrové informační matice $F_n(\boldsymbol{\beta})$ a $\mathcal{I}_n(\boldsymbol{\beta})$. Průběh algoritmů se ovšem zásluhou různých počátečních hodnot může lišit.

Povšimněme si, že (1.4.5) odpovídá vzorci ve vážené lineární regresi. V každé iteraci se vlastně stanoví nový odhad parametru $\boldsymbol{\beta}$ pomocí vážené metody nejmenších čtverců s linearizovanou odezvou, maticí plánu $\mathbf{X}$ a příslušnými hodnotami vah. Odtud také pochází samotný název. Existence této interpretace je velice výhodná, jelikož jsou známé velmi efektivní numerické algoritmy pro výpočet aktualizované hodnoty $\boldsymbol{\beta}$. Navíc pro počáteční odhad nutně nepotřebujeme znát hodnotu $\hat{\boldsymbol{\beta}}^{(0)}$, ale stačí nám $\hat{\boldsymbol{\mu}}^{(0)}$, což je také výhodné.

1.5 Asymptotické vlastnosti GLM

V této sekci bylo čerpáno především z dizertační práce [8] , ale také z článku [6]. Článek diskutuje různé varianty předpokladů pro splnění asymptotické normality a konzistence maximálně věrohodného odhadu, a to jak při náhodných tak nenáhodných regresorech. Formulujeme pro úplnost alespoň jeho nejdůležitější závěry.

V následující větě bude $\boldsymbol{\beta}_0$ značit skutečnou hodnotu parametru $\boldsymbol{\beta} \in \mathbf{B} \subseteq \mathbb{R}^p$ a $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n(\mathbf{Y})$ náhodný vektor, který je pro realizaci $\mathbf{Y}$ roven maximálně věrohodnému odhadu, pokud existuje. Fisherovu informační matici $\mathcal{I}_n(\boldsymbol{\beta})$ můžeme ze spektrálního rozkladu vyjádřit jako

$$\mathcal{I}_n(\boldsymbol{\beta}) = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}^T = \mathbf{C}\boldsymbol{\Lambda}^{1/2}\mathbf{C}^T\mathbf{C}\boldsymbol{\Lambda}^{1/2}\mathbf{C}^T = \mathcal{I}_n^{1/2}(\boldsymbol{\beta})\mathcal{I}_n^{1/2}(\boldsymbol{\beta}),$$

kde $\mathbf{C}$ je matice s ortonormálními vlastními vektory $\mathcal{I}_n(\boldsymbol{\beta})$ ve sloupcích, Λ diagonální matice vlastních čísel a $\boldsymbol{\Lambda}^{1/2}$ diagonální matice odmocnin vlastních čísel. Pro $\boldsymbol{\beta} \in \mathbf{B}$ a libovolné $\delta > 0$ označme

$$\bar{\mathcal{O}}(\delta) = \{\boldsymbol{\beta} : \| \mathcal{I}_n^{1/2}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \| \leq \delta\}.$$

Symbol $\| \cdot \|$ zde chápeme jako euklidovskou normu vektoru. Nakonec položme

$$\mathbf{V}_n(\beta) = \mathcal{I}_n^{-1/2}(\boldsymbol{\beta}_0)\mathcal{I}_n(\boldsymbol{\beta})\mathcal{I}_n^{-1/2}(\boldsymbol{\beta}_0).$$

V nadcházející větě budeme pracovat s následujícími výroky:

(P1). g je kanonická linková funkce.

(P2). $\lim_{n\to\infty}\lambda_{min}(\mathcal{I}_n(\boldsymbol{\beta}_0)) = \infty$, kde $\lambda_{min}(\mathcal{I}_n(\boldsymbol{\beta}_0))$ je nejmenší vlastní číslo $\mathcal{I}_n(\boldsymbol{\beta}_0)$.

(P3). Pro každé $\delta > 0$ existuje n_1 a $c > 0$ takové, že pro všechna $n > n_1$ a $\boldsymbol{\beta} \in \bar{\mathcal{O}}(\delta)$ je matice $\mathcal{I}_n(\boldsymbol{\beta}) - c\mathcal{I}_n(\boldsymbol{\beta}_0)$ pozitivně semidefinitní.

(P4). Pro matici $\mathbf{V}_n(\beta) = \mathcal{I}_n^{-1/2}(\boldsymbol{\beta}_0)\mathcal{I}_n(\boldsymbol{\beta})\mathcal{I}_n^{-1/2}(\boldsymbol{\beta}_0)$ a každé $\delta > 0$ platí

$$\lim_{n\to\infty}\max_{\boldsymbol{\beta}\in\bar{\mathcal{O}}(\delta)} \parallel \mathbf{V}_n(\boldsymbol{\beta}) - \mathcal{I}_p \parallel = 0,$$

přičemž symbol $\| \cdot \|$ zde chápeme jako normu matice kompatibilní (souhlasnou) s Euklidovskou normou.

(P5). Posloupnost $\{\mathbf{x}_n^T\}_{n=1}^{\infty}$ leží celá v kompaktní množině $\mathcal{X}$ takové, že $\mathbf{x}^T\boldsymbol{\beta} \in \Theta$ pro každé $\boldsymbol{\beta} \in \mathbf{B}$ a $\mathbf{x}^T \in \mathcal{X}$.

(P6). $\lim_{n\to\infty}\lambda_{min}(\sum_{i=1}^{n}(\frac{\tilde{w}_i}{\varphi})^2\mathbf{x}_i\mathbf{x}_i^T) = \infty$, kde $\lambda_{min}(\sum_{i=1}^{n}(\frac{\tilde{w}_i}{\varphi})^2\mathbf{x}_i\mathbf{x}_i^T)$ je nejmenší vlastní číslo $\sum_{i=1}^{n}(\frac{\tilde{w}_i}{\varphi})^2\mathbf{x}_i\mathbf{x}_i^T$.

Věta 1.6. *Nechť* $\mathbf{Y} \sim GLM(f, g, \mathbf{X})$, $\mathbf{X}$ *má lineárně nezávislé sloupce,* b *je třikrát spojitě diferencovatelná (*$b \in C^3(\Theta)$*) a* $\mathbf{B} \subseteq \mathbb{R}^p$ *je otevřená konvexní množina. Nechť navíc pro každé* $\boldsymbol{\beta} \in \mathbf{B}$ *platí* $\mathbf{x}_i^T\boldsymbol{\beta} \in g(\mu(\Theta)), i = 1, \ldots, n$, *kde* $g(\mu(\Theta)) = \{g(b'(\theta)) : \theta \in \Theta\}$.

(1) Jsou-li splněny předpoklady (P1),(P2) a (P3), pak existuje posloupnost náhodných vektorů $\{\hat{\boldsymbol{\beta}}_n\}_{n=1}^{\infty}$ *taková, že*

(Z1). $\hat{\boldsymbol{\beta}}_n$ *je asymptoticky řešením věrohodnostních rovnic, tj.,*

$$\lim_{n\to\infty}P\{U(\hat{\boldsymbol{\beta}}_n) = 0\} = 1,$$

(Z2). $\hat{\boldsymbol{\beta}}_n$ *je zároveň konzistentním odhadem* $\boldsymbol{\beta}_0$, *tj.*

$$\lim_{n\to\infty}P\{\parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \parallel < \varepsilon\} = 1.$$

(2) Pokud platí (P1),(P2) a (P4), můžeme k závěrům (Z1) a (Z2) přidat

(Z3). $\{\hat{\boldsymbol{\beta}}_n\}$ *je asymptoticky normální,tj.*

$$\mathcal{I}_n^{1/2}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, I_p),$$

kde I_p *je jednotková matice řádu* p.

(Z4). Skórový vektor $U(\boldsymbol{\beta}_0)$ *je také asymptoticky normální při* $n \to \infty$, *tj.*

$$\mathcal{I}_n^{-1/2}(\boldsymbol{\beta}_0)U(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, I_p).$$

(3) Výroky (P1), (P5) a (P6) také implikují (Z1), (Z2), (Z3) a (Z4).

Důkaz. Viz dizertační práce [8], věty 3.15, 3.16 a důsledek 3.19, případně je také možno nahlédnout do článku [6]. □

Podmínka (P6) je nutná a postačující pro konzistenci odhadu v klasické lineární regresi. Při pevné hodnotě počtu regresorů nebývá v praxi s jejím splněním (dle [13] podsekce 6.2.4) jakýkoli problém. Předpoklad (P3) lze interpretovat jako požadavek, aby nedocházelo na okolí skutečné hodnoty parametru ke ztrátě informace, a (P4) jako požadavek dostatečné spojitosti Fisherovy informační matice. Oba tyto předpoklady lze modifikovat i pro případy s libovolnou linkovou funkcí (viz [6] str. 360-361). Nakonec (P5), který vyžaduje, aby regresory pocházely z kompaktního prostoru, lze snadno ověřit. Modifikaci předpokladů pro linkové funkce, které nejsou kanonické, lze také nalézt v [8] (str. 48, poznámka 3.22).

Poznámka. Budeme-li se v další kapitolách zmiňovat o asymptotických vlastnostech, implicitně předpokládáme splnění předpokladů této věty.

Kapitola 2

Aplikace a verifikace

V následující kapitolách bylo primárně čerpáno z [4]. Blíže se zde zaměříme na modelování počtu pojistných událostí vzniklých v souvislosti s danou pojistnou smlouvou za dané období. Odhad četností škod tvoří jednu ze zásadních složek pro stanovení pojistného. Pojistný matematik se snaží především co nejlépe vyčíslit množství prostředků odpovídajících očekávané výši všech škod, kterou označujeme jako netto pojistné. Standardně se předpokládá, že jednotlivé výše škod jsou vzájemně nezávislé a stejně rozdělené a zároveň nezávisí na svém počtu, proto bývá celková výše škod vzniklá na dané smlouvě modelována pomocí vhodného složeného rozdělení. Netto pojistné poté jednoduše odpovídá součinu střední hodnoty výše a střední hodnoty počtu škod.

Základní metodou pro posouzení rizikovosti pojistné smlouvy je odhad očekávané škodní frekvence, tj. počtu pojistných událostí v případě, kdy by smlouva byla v platnosti po celé sledované období odpovídající zpravidla jednomu roku. Pro pojištění odpovědnosti z provozu motorových vozidel probíhá odhad očekávané škodní frekvence obvykle ve dvou fázích.

První fázi se budeme podrobněji věnovat po zbytek této práce. Mohli bychom ji nazvat apriorní, jelikož je založena pouze na předem známých údajích uvedených v pojistné smlouvě. Takovýmto údajům říkáme rizikové charakteristiky smlouvy a může mezi ně patřit například věk pojistníka, pohlaví, povolání, zdvihový objem motoru vozidla, velikost bydliště pojištěného, rodinný stav či frekvence placení pojistného. Příhodným nástrojem pro tuto fázi jsou právě zobecněné lineární modely, které zahrnují informace o rizikových charakteristikách prostřednictvím nezávisle proměnných.

V druhé aposteriorní fázi se odhad škodní frekvence upravuje na základě historického škodního průběhu nebo zohledněním postavení pojištěného v rámci systému bonus-malus. Pojišťovny používají systém k motivaci řidičů, aby jezdili s co nejmenším počtem nehod. Zodpovědní řidiči jezdící bez nehody jsou v rámci žebříčku posouváni do kategorií s vyšší slevou na pojistném, jsou oceněni bonusem. Naopak řidiči, kteří způsobí nehodu, jsou penalizováni tzv. malusem, tj. jsou posunuti opačným směrem a naopak platí vyšší pojistné. Podrobněji se touto problematikou zabývá kniha [4] v kapitole 4.

2.1 Aplikace GLM v pojišťovnictví

2.1.1 Poissonovská regrese

V minulosti již vzniklo více zdařilých pokusů o použití různých pravděpodobnostních modelů pro odhad frekvencí pojistných událostí. Obecně se všechny opírají o předpoklad, že počet škod způsobených pojištěným během sledovaného období se řídí Poissonovým rozdělení, které si proto drží prominentní postavení v této oblasti.

Základním nejjednodušším používaným zobecněným lineárním modelem je tzv. poissonovská regrese. Náhodná veličina N_i reprezentuje počet škod nastalých na smlouvě i, která je v platnosti po dobu d_i ve sledovaném období, které zpravidla odpovídá jednomu roku. Informace obsažené v rizikových charakteristiky budou reprezentovány pomocí $p-1$ nezávisle proměnných vektorem $(x_{i1}, \ldots, x_{ip-1})^T$. Na model jsou kladeny následující tři základní požadavky:

1. Předpis pro střední hodnotu má tvar

$$\mathrm{E}\, N_i = d_i \exp(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}), \qquad i = 1, 2, \ldots, n, \tag{2.1.1}$$

   kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ odpovídá vektoru neznámých regresních koeficientů délky p, které máme za cíl odhadnout, a n počtu smluv.

2. Vysvětlované proměnné $N_1, \ldots, N_n$ jsou vzájemně nezávislé.

3. Náhodné veličiny $N_1, \ldots, N_n$ mají navíc Poissonovo rozdělení, tj.

$$N_i \sim Po(d_i \exp(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij})), \qquad i = 1, 2, \ldots, n.$$

V souladu se značením v předchozí kapitole je $N_i \sim Po(d_i \exp(\mathbf{x}_i^T \boldsymbol{\beta}))$, kde $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip-1})$. Parametr β_0 nazýváme intercept a bez ohledu na jeho výskyt budeme v modelu nadále konzistentně předpokládat právě p neznámých regresních koeficientů. Poznamenejme, že obecněji můžeme výskyt pojistných událostí na smlouvě i pokládat za realizaci homogenního Poissonova procesu, neboť $\mathrm{E}\, N_i$ je součinem střední hodnoty (roční) škodní frekvence $\exp(\mathbf{x}_i^T \boldsymbol{\beta})$ a doby d_i.

2.1.2 Expozice v riziku

Dobu d_i nazýváme expozice v riziku. Přirozeně většina smluv zůstává v platnosti po celé sledované období jednoho roku a má tedy expozici rovnu jedné. Ve zbylých případech je $d_i < 1$, což nastává nejen, pokud dojde ke zrušení či vzniku smlouvy v průběhu sledovaného roku, ale také v případě, kdy se změní hodnota některého z regresorů. Z hlediska modelu můžeme chápat jakoukoli takovou změnu jako zánik staré a vznik nové smlouvy. Povšimněme si, že poté považujeme škodní průběhy pojištěného v různých obdobích za nezávislé, což určitě není realistické. Problém

korelovanosti dat lze vyřešit užitím tzv. GEE metody, o které bude blíže pojednáno v závěru příští kapitoly.

Důležitým často používaným pojmem je offset. Povšimněme si, že výše uvedený výraz (2.1.1) lze také zapsat ve tvaru

$$\mathrm{E}\, N_i = d_i \exp(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}) = \exp(\ln d_i + \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}),$$

kde se v exponentu navíc vyskytuje hodnota $\ln d_i$. Tuto hodnotu označujeme anglickým termínem offset. Aniž by došlo k zásadním změnám v poznatcích uvedených v předchozí kapitole, můžeme symbol $\eta_i = \log \mathrm{E}\, N_i$ (viz (1.2.2)) také definovat jako součet lineárního prediktoru a offsetu, tj. $\eta_i = \ln d_i + \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$. Existuje ovšem ještě další možnost, jak zařadit do modelu informaci o expozicích. Pracujeme-li s ročními škodními frekvencemi $(\mathrm{E}\, N_i/d_i, i = 1, \ldots, n)$ jakožto nezávisle proměnnými, pak lze d_i chápat jako apriorní váhu (prior weight) $\tilde{w}_i$ ve vyjádření (1.2.1), tj. $d_i = \tilde{w}_i$. Z (1.3.6) plyne, že odhad regresních koeficientů je v obou případech identický. V dalších úvahách budeme předpokládat, že váhy $\tilde{w}_1, \ldots, \tilde{w}_n$ jsou rovny jedné a expozice jsou začleněny do modelu pomocí offsetu.

2.1.3 Interpretace modelu

Ve výše uvedené formulaci modelu byla pro vazbu mezi střední hodnotou počtu škod a lineárním prediktorem použita logaritmická linková funkce, která je zároveň pro Poissonovo rozdělení linkem kanonickým (viz podsekce 1.1.1). To poskytuje odhadu kromě příhodných statistických vlastností také velmi rozumnou a přímočarou interpretaci vlivu nezávisle proměnné na očekávanou škodní frekvenci, a proto se v praxi pracuje výhradně s touto vazbou.

Rizikové charakteristiky, které mají charakter spojitých či ordinálních proměnných bývají obvykle kategorizovány.[1] Důvodem je, že zpravidla apriorně neznáme vhodné transformace těchto proměnných, jež by odpovídaly funkčnímu předpisu modelu. Problém se navíc komplikuje, uvědomíme-li si, že závislostní struktura může býti poměrně komplikovaná a přidáním či odebráním některého z regresorů se vliv ostatních může změnit, a tedy původně použité transformace nemusí být vůbec smysluplné.

Příklad. Vysvětleme na jednoduchém příkladě, jak může kategorizace a následná interpretace vypadat. Mějme k dispozici portfolio smluv povinného ručení a zároveň informace o třech rizikových charakteristikách, kterými jsou: zdvihový objem motoru, věk pojistníka a pohlaví. Ty do modelu zahrneme pomocí dummy proměnných [2]

[1] Ordinální proměnné nabývají konečně mnoha hodnot a lze je uspořádat na základě kvalitativní charakteristiky, kterou vystihují.

[2] Tj. proměnných nabývající pouze hodnot 0 nebo 1.

$$x_{i1} = \begin{cases} 1 & \text{pokud je objem motoru menší než } 1400[cm^3] \\ 0 & \text{jinak} \end{cases},$$

$$x_{i2} = \begin{cases} 1 & \text{pokud je objem motoru 1400 až } 2000[cm^3] \\ 0 & \text{jinak} \end{cases},$$

$$x_{i3} = \begin{cases} 1 & \text{pokud je objem motoru větší než } 2000[cm^3] \\ 0 & \text{jinak} \end{cases},$$

$$x_{i4} = \begin{cases} 1 & \text{pokud je řidič vozidla muž} \\ 0 & \text{jinak} \end{cases},$$

$$x_{i5} = \begin{cases} 1 & \text{pokud je věk menší než 26 let} \\ 0 & \text{jinak} \end{cases},$$

$$x_{i6} = \begin{cases} 1 & \text{pokud je věk 26 až 30 let} \\ 0 & \text{jinak} \end{cases},$$

$$x_{i7} = \begin{cases} 1 & \text{pokud je věk vyšší než 60 let} \\ 0 & \text{jinak} \end{cases}.$$

Řidiči mezi 31 a 60 lety tvoří referenční skupinu pro rizikovou charakteristiku věk a obdobně je tomu u žen v rámci pohlaví. Povšimněme si, že do modelu nebyl zahrnut intercept. Vztah pro očekávanou škodní frekvenci $(\mathrm{E}\, N_i)/d_i$ by pak měl tvar:

$$\exp(\beta_1 x_{i1} + \cdots + \beta_7 x_{i7}) = \prod_{j|x_{ij}=1} \exp(\beta_j).$$

Hodnoty $\exp(\beta_1), \exp(\beta_2)$ a $\exp(\beta_3)$ jsou očekávanými škodními frekvencemi skupiny referenčních kategorií (tj. žen ve věku 31 až 60 let) pro různé objemy motorů či obecněji tarifní skupiny. Smlouvy povinného ručení totiž bývají v pojišťovnách zařazovány do tarifních skupin právě na základě objemu, typu vozidla či síly motoru. Pojem by neměl být zaměňován s termínem tarifní třída, kterou určuje teprve celý vektor $(x_{i1}, \ldots, x_{i7})^T$. Pro $j = 4, \ldots, 7$, určují vyjádření $(\exp(\beta_j) - 1) \cdot 100$, o kolik procent se zvýší (resp. sníží) škodní frekvence vlivem j-té proměnné vůči referenční kategorii. Například odhad β_5 interpretujeme následovně: Mladí řidiči (resp. řidičky) jsou ve všech tarifních skupinách rizikovější o $(\exp(\beta_5) - 1) \cdot 100\%$ než muži (resp. ženy) ve věku 31 až 60 let.

Poznámka. Situace je vlastně velmi podobná jako v případě klasického lineárního modelu, až na fakt, že díky logaritmickému linku se změní aditivní vliv jednotlivých nezávisle proměnných na multiplikativní. Pro úplnost poznamenejme, že zařazením interceptu do modelu nezískáme nic jiného než jinou interpretaci, navíc je přitom nutné vyřadit jeden z regresorů identifikujících tarifní skupinu, která se stane novou referenční kategorií.

Často se stává, že rizikové charakteristiky mezi sebou vzájemně interagují a nelze je do modelu zahrnout odděleně. Typickým příkladem může být právě věk a pohlaví (viz [4], sekce 2.2, obr. 2.7). Mladí řidiči vykazují výrazně vyšší škodní frekvence než mladé řidičky, naopak ale muži ve středním věku vykazují spíše nižší frekvence než ženy ve stejném věku. Do modelu jsou poté zahrnuty dummy proměnné pro každou kombinaci kategorií interagujících proměnných s výjimkou kombinace referenčních kategorií. Ve výše uvedeném případě by tak vznikly nové binární regresory identifikující následující skupiny: muži 18-24, muži 25-30, muži 31-60, muži nad 60, ženy 18-24, ženy 25-30, ženy nad 60. Na ženy ve věku 31-60 let by tak zbyla referenční kategorie a výsledný model by obsahoval deset dummy proměnných.

2.1.4 Interpretace věrohodnostních rovnic

Nechť k_i označuje počet nahlášených pojistných událostí na smlouvě i. Věrohodnostní funkce pro tato pozorování má tvar

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} P[N_i = k_i] = \prod_{i=1}^{n} \exp(-\lambda_i)\frac{\lambda_i^{k_i}}{k_i!},$$

kde

$$\lambda_i = d_i \exp(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}), \quad i = 1, \ldots, n.$$

Logaritmická věrohodnostní funkce je dána předpisem

$$L(\boldsymbol{\beta}) = \ln \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} (-\ln k_i! + k_i \ln \lambda_i - \lambda_i),$$

a pro příslušné věrohodnostní rovnice platí:

$$\frac{\partial}{\partial \beta_0} L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\left(k_i \frac{\partial}{\partial \beta_0} \ln \lambda_i \right) - \frac{\partial}{\partial \beta_0} \lambda_i \right) = \sum_{i=1}^{n} (k_i - \lambda_i) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} k_i = \sum_{i=1}^{n} \lambda_i, \tag{2.1.2}$$

$$\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\left(k_i \frac{\partial}{\partial \beta_j} \ln \lambda_i \right) - \frac{\partial}{\partial \beta_j} \lambda_i \right) = \sum_{i=1}^{n} (k_i x_{ij} - \lambda_i x_{ij}) = 0 \tag{2.1.3}$$

$$\Leftrightarrow \sum_{i=1}^{n} k_i x_{ij} = \sum_{i=1}^{n} \lambda_i x_{ij}, \quad j = 1, \ldots, p-1, \tag{2.1.4}$$

což zároveň odpovídá (1.3.6) při $\tilde{w}_i = 1, i = 1, \ldots, n$. První rovnice nám říká, že součet odhadů středních hodnot počtů škod $\sum_{i=1}^{n} \hat{\lambda}_i$ je roven celkovému pozorovanému počtu $\sum_{i=1}^{n} k_i$. To zřejmě platí pouze, vyskytuje-li se v lineárním prediktoru intercept. Podobnou interpretaci má i druhá rovnice, pokud jsou regresory dummy

proměnné. Například ve výše uvedeném příkladě (bez interakce) by pro $j = 4$ vztah přešel v

$$\sum_{muži} k_i = \sum_{muži} \hat{\lambda}_i,$$

tj. součet vyrovnaných hodnot týkajících se mužů je roven celkovému počtu pojistných událostí způsobených muži.

2.2 Verifikace

V této sekci budou představeny některé verifikační techniky určené pro model poissonovské regrese, přičemž níže uvedené poznatky povětšinou vycházejí z výsledků platných i v širším kontextu zobecněných lineárních modelů (viz např. [9] sekce 5.8 a [5] sekce 3.3). Více obecné přístupy (viz sekce 3.1) někdy vyžadují nahrazení disperzního parametru jeho odhadem, a proto budeme v některých níže uvedených vyjádřeních ponechávat disperzní parametr, i když je ve skutečnosti dle klasické poissonovské regrese roven jedné. Připomeňme, že nadále předpokládáme váhy $\tilde{w}_1, \ldots, \tilde{w}_n$ rovny jedné.

Na konci přílohy A jsou popsány tři testy na základě kterých lze rozhodnout, zda daná podmnožina skupiny regresorů určujících podmodel nedokáže vysvětlit nezávisle proměnnou stejně kvalitně jako celá skupina, tj. nulová hypotéza odpovídá platnosti podmodelu. Test poměrem věrohodností, skórový i Waldův se užívají i v případě zobecněných lineárních modelů. Ukažme si nyní podrobněji jak vypadají intervaly spolehlivosti a testy významnosti parametrů.

2.2.1 Intervaly spolehlivosti a testy hypotéz

Běžně bývá v literatuře zabývající se modelováním četností pojistných událostí pomocí regresních metod na prvním místě zmiňována Waldova testová statistika (viz (A.3.2)). Důvod její časté aplikace pro konstrukce intervalů spolehlivosti a testování hypotéz se skrývá v její nízké citlivosti na správnou specifikaci modelu. Software obvykle vyžaduje pro testování nulovosti parametrů zadání tzv. matice hypotézy, označme ji $\mathbf{C}$, a ověřuje, zda platí hypotéza $\mathbf{C}\boldsymbol{\beta} = 0$, přičemž předpokládáme, že $\mathbf{C}$ je matice plné hodnosti typu $k \times p$. Jelikož maximálně věrohodný odhad $\hat{\boldsymbol{\beta}}$ má (dle věty 1.6) pro velká n přibližně normální rozdělení, tj. $\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \mathcal{I}_n^{-1})$, platí pro $\mathbf{C}\hat{\boldsymbol{\beta}}$ (za nulové hypotézy) vztah

$$\mathbf{C}\hat{\boldsymbol{\beta}} \approx N(0, \mathbf{C}\mathcal{I}_n^{-1}\mathbf{C}^T), \tag{2.2.1}$$

kde $\mathcal{I}_n = \frac{1}{\varphi}\sum_{i=1}^n \tilde{w}_i w(\mu_i)\mathbf{x}_i\mathbf{x}_i^T = \frac{1}{\varphi}\mathbf{X}^T\mathbf{W}\mathbf{X}$ (viz (1.4.4)). Z toho dále plyne, že Waldova statistika

$$W = (\mathbf{C}\hat{\boldsymbol{\beta}})^T(\varphi\mathbf{C}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{C}^T)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}) \approx \chi_k^2, \tag{2.2.2}$$

má přibližně χ_k^2 rozdělení, kde k označuje počet řádků matice $\mathbf{C}$, jež odpovídají jednotlivým omezením,$\hat{\mathbf{W}} = diag(w_1 w(\hat{\mu}_1), \ldots, w_n w(\hat{\mu}_n))$. Speciálně v případě poissonovské regrese je $\varphi = 1$, $\mathcal{I}_n = \sum_{i=1}^n \lambda_i \mathbf{x}_i\mathbf{x}_i^T$ a $\hat{\mathbf{W}} = diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$. Toto vyjádření

kromě testování významnosti skupin regresorů umožňuje také například ověřit rovnosti dvojic parametrů. Poznamenejme ještě, že v následující kapitole (sekce 3.1) je naznačeno, jak modifikovat tento přístup pro případ, kdy si nejsme jisti splněním některým ze základních požadavků kladených na model.

Například hypotézu $\beta_j = 0$ ověříme volbou $\mathbf{C}$ jakožto řádkového vektoru, který obsahuje jedinou nenulovou složku na pozici j, jež je rovna jedné. Výše uvedený výraz (2.2.2) přejde v

$$\frac{\hat{\beta}_j^2}{\varphi v_j} = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2_{\hat{\beta}_j}} \approx \chi^2_1, \tag{2.2.3}$$

kde v_j je j-tým diagonální prvkem matice $(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}$, $\hat{\sigma}^2_{\hat{\beta}_j}$ odhad rozptylu $\hat{\beta}_j$ (tj. j-tý diagonální prvek $\mathcal{I}_n^{-1}$) a φ disperzní parametr. Nulovost parametru zamítáme při $W > \chi^2_{1,1-\alpha}$, kde $\chi^2_{1,1-\alpha}$ je $(1-\alpha)$ kvantil rozdělení χ^2_1. Hypotézu $\beta_i = \beta_j$ otestujeme volbou

$$\mathbf{C} = (0, \dots 0, \underbrace{1}_{i}, 0, \dots 0, \underbrace{1}_{j}, 0, \dots, 0).$$

Interval spolehlivosti pro β_j lze snadno zkonstruovat s využitím (2.2.1). Na hladině spolehlivosti $1-\alpha$ jej můžeme zapsat ve tvaru

$$\left[\hat{\beta}_j - u_{\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\beta}_j}}, \hat{\beta}_j + u_{\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\beta}_j}}\right],$$

kde $u_{\alpha/2}$ je $(\alpha/2)\cdot 100\%$-ní kvantil normálního rozdělení. Pokud platí $\hat{\boldsymbol{\beta}} \sim N(\beta, \mathcal{I}_n^{-1})$, pak $\mathbf{x}_i^T\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, \mathbf{x}_i^T\mathcal{I}_n^{-1}\mathbf{x}_i)$, z čehož dostáváme díky ryzí monotonii exponenciální funkce pro škodní frekvenci interval spolehlivosti tvaru

$$\left[\exp\left(\mathbf{x}_i^T\boldsymbol{\beta} - u_{\alpha/2}\sqrt{\mathbf{x}_i^T\mathcal{I}_n^{-1}\mathbf{x}_i}\right), \exp\left(\mathbf{x}_i^T\boldsymbol{\beta} + u_{\alpha/2}\sqrt{\mathbf{x}_i^T\mathcal{I}_n^{-1}\mathbf{x}_i}\right)\right].$$

Samozřejmě v praxi se $\boldsymbol{\beta}$ nahrazuje $\hat{\boldsymbol{\beta}}$ a $\mathcal{I}_n$ odhadem $\sum_{i=1}^n \hat{\lambda}_i\mathbf{x}_i\mathbf{x}_i^T$ nebo případně jiným robustnějším odhadem (viz následující kapitola sekce 3.1).

Pokud není důvod pochybovat o zvoleném rozdělení pro zobecněný lineární model, doporučuje se (dle [9]) z důvodu větší síly použít pro verifikaci test poměrem věrohodností. Nevhodnost poissonovské regrese se typicky projeví v odhadu disperzního parametru, který by měl být přibližně roven jedné a o němž bude ještě dále pojednáno (viz sekce 3.1). Nechť má vektor $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})^T$ po vhodném přeuspořádání svých složek tvar

$$\begin{pmatrix} \boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)} \end{pmatrix},$$

kde $\boldsymbol{\beta}_{(1)}$ je délky q $(0 < q < p)$, pak test nulovosti parametrů vypadá následovně:

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_{(2)} = 0, \text{ tj. platí podmodel,} \\ H_1 &: \boldsymbol{\beta}_{(2)} \neq 0, \text{ tj. platí model.} \end{aligned} \tag{2.2.4}$$

Nechť jsou $\hat{\boldsymbol{\beta}}_0$ a $\hat{\boldsymbol{\beta}}_1$ po řadě maximálně věrohodnými odhady v podmodelu (tj. za platnosti H_0) a modelu, pak má testová statistika

$$2[L(\hat{\boldsymbol{\beta}}_1) - L(\hat{\boldsymbol{\beta}}_0)] \tag{2.2.5}$$

za nulové hypotézy (dle [4] podsekce 2.3.12) asymptoticky χ^2_{p-q} rozdělení. Z čehož úpravou dostáváme asymptotický $(1-\alpha)\cdot 100\%$ interval spolehlivosti pro každé $\beta_j, j = 0, \ldots, p-1$, jako množinu tvaru

$$\{\beta_j : L_{\beta_j} \geq L(\hat{\boldsymbol{\beta}}_1) - \frac{1}{2}\chi^2_{1,1-\alpha}\},$$

kde L_{β_j} je hodnota logaritmické věrohodnostní funkce spočtená na základě maximálně věrohodného odhadu při pevném β_j, tj. dle předchozího značení v podmodelu s $\boldsymbol{\beta}_{(2)} = \beta_j$ a $q = p-1$.

2.2.2 Deviance

Mějme věrohodnostní funkci $\tilde{\ell}(\boldsymbol{\lambda}) = \prod_{i=n}^{n} \exp(-\lambda_i)\frac{\lambda_i^{k_i}}{k_i!}$, kde $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$. Je snadné ověřit, že $\tilde{\ell}(\boldsymbol{\lambda})$ nabývá svého maxima při $\boldsymbol{\lambda} = \mathbf{k} = (k_1, \ldots, k_n)^T$. Proto věrohodnostní funkce dosahuje své nejvyšší hodnoty právě, když odhady středních hodnot počtů škodných událostí (vyrovnané hodnoty $\hat{\lambda}_i$) jsou rovny svým pozorovaným hodnotám k_i. Toto je charakteristické pro tzv. saturovaný model, jež obsahuje shodný počet parametrů a pozorování.

Nechť $\hat{\boldsymbol{\beta}}$ je maximálně věrohodný odhad regresních koeficientů a $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_n)^T$ $= (d_1 \exp(\mathbf{x}_1^T\hat{\boldsymbol{\beta}}), \ldots, d_n \exp(\mathbf{x}_n^T\hat{\boldsymbol{\beta}}))$. Deviance $D(\mathbf{k}, \hat{\boldsymbol{\beta}})$ je definována jako dvojnásobek rozdílu logaritmických věrohodnostních funkcí saturovaného a současného modelu, tj.

$$\begin{aligned} D(\mathbf{k}, \hat{\boldsymbol{\beta}}) &= -2\ln \frac{\tilde{\ell}(\hat{\boldsymbol{\lambda}})}{\tilde{\ell}(\mathbf{k})} = 2(\tilde{\mathrm{L}}(\mathbf{k}) - \tilde{\mathrm{L}}(\hat{\boldsymbol{\lambda}})) \\ &= 2\ln\left(\prod_{i=n}^{n} \exp(-k_i)\frac{k_i^{k_i}}{k_i!}\right) - 2\ln\left(\prod_{i=n}^{n} \exp(-\hat{\lambda}_i)\frac{\hat{\lambda}_i^{k_i}}{k_i!}\right) \\ &= 2\sum_{i=1}^{n}\left(k_i \ln\frac{k_i}{\hat{\lambda}_i} - (k_i - \hat{\lambda}_i)\right), \end{aligned}$$

kde $y\ln y = 0$ pro $y = 0$ dle konvence. Jde tedy o jakousi odchylku či vzdálenost současného modelu od nejlepší možné „dosažitelné“ hranice, kterou představuje model saturovaný.

Test poměrem věrohodností se také někdy nazývá devianční. Deviance má totiž úzkou souvislost s jeho testovou statistikou, neboť ta přesně odpovídá rozdílu deviancí. Pro dvojici (model, podmodel) specifikovanou výše (viz (2.2.4)) platí:

$$D(\mathbf{k}, \hat{\boldsymbol{\beta}}_0) - D(\mathbf{k}, \hat{\boldsymbol{\beta}}_1) = 2[\tilde{L}(\boldsymbol{k}) - L(\hat{\boldsymbol{\beta}}_0)] - 2[\tilde{L}(\boldsymbol{k}) - L(\hat{\boldsymbol{\beta}}_1)] = 2[L(\hat{\boldsymbol{\beta}}_1) - L(\hat{\boldsymbol{\beta}}_0)].$$

Poznamenejme, že deviance $D(\mathbf{k}, \hat{\boldsymbol{\beta}}_0)$ nemá obecně χ^2_{n-q} rozdělení. Asymptotické vlastnosti maximálně věrohodného odhadu se totiž neuplatní, jelikož se zvyšujícím počtem pozorování roste zároveň množství parametrů.

Při verifikaci také někdy pracujeme s modelem zkonstruovaným na základě dat seskupených dle tarifních tříd, o nichž budeme dále předpokládat, že jsou jednoznačně určeny hodnotami lineárního prediktoru. Odhad $\boldsymbol{\beta}$ zůstává i v takovém případě identický, ale dojde k posunu věrohodnostní funkce, díky kterému získá deviance přibližně χ^2 rozdělení.

Nechť $s_1, \ldots, s_q$ označují q možných hodnot lineárního prediktoru ($q < \infty$), který pro i-tou smlouvu nabývá $score_i$. Definujme

$$d_{\bullet j} = \sum_{i|score_i = s_j} d_i \quad \text{a} \quad k_{\bullet j} = \sum_{i|score_i = s_j} k_i \quad \text{pro } j = 1, \ldots, q.$$

Součty $d_{\bullet j}$ a $k_{\bullet j}$ jsou charakteristikami j-té tarifní třídy, mající význam celkové expozice a celkového počtu škod. Úpravou věrohodnostní funkce dostáváme:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \prod_{j=1}^{q} \prod_{i|score_i = s_j} \exp(-\lambda_i) \frac{\lambda_i^{k_i}}{k_i!} \\
&= konst. \cdot \prod_{j=1}^{q} \exp\Big(-\sum_{i|score_i = s_j} \lambda_i\Big) (\exp(s_j) d_{\bullet j})^{k_{\bullet j}} \\
&= konst. \cdot \prod_{j=1}^{q} \exp\Big(-\exp(s_j) \sum_{i|score_i = s_j} d_i\Big) (\exp(s_j) d_{\bullet j})^{k_{\bullet j}} \\
&= konst. \cdot \underbrace{\prod_{j=1}^{q} \exp(-\exp(s_j) d_{\bullet j}) \frac{(\exp(s_j) d_{\bullet j})^{k_{\bullet j}}}{k_{\bullet j}!}}_{\ell_{seskup}(\boldsymbol{\beta})}. \qquad (2.2.6)
\end{aligned}$$

Vyjádření $\ell_{seskup}(\boldsymbol{\beta})$ je zřejmě věrohodnostní funkcí modelu se seskupenými daty. Ukázali jsme tedy, že $\ell(\boldsymbol{\beta})$ a $\ell_{seskup}(\boldsymbol{\beta})$ se liší pouze o násobek konstantou, což je zároveň důvod, proč jsou odhady $\boldsymbol{\beta}$ v obou případech identické. Odhad Fisherovy informační matice zůstává také stejný, neboť platí:

$$\mathcal{I}_n(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T d_i \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i) = \sum_{j=1}^{q} \mathbf{x}_j^{seskup} (\mathbf{x}_j^{seskup})^T d_{\bullet j} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_j^{seskup}),$$

kde $\mathbf{x}_j^{seskup}$ je vektor regresorů odpovídající j-té tarifní třídě, tj. $\mathbf{x}_j^{seskup} = \mathbf{x}_i$, pokud i-tá smlouva spadá do j-té tarifní třídy. Deviance má asymptoticky χ^2_{q-p} při $d_{\bullet j} \to \infty$ pro všechna $j = 1, \ldots, q$, kde q označuje počet tarifních tříd a p fixní počet regresních koeficientů (blíže viz kniha [4] podsekce 2.3.11 nebo článek [14]).

2.2.3 Rezidua

Rezidua lze v souvislosti se zobecněnými lineárními modely využít k posouzení kvality modelu, detekci odlehlých pozorování a ověření předpokladu o rozptylu. V literatuře se můžeme setkat s třemi typy reziduí. Jsou to: Pearsonova, Anscombova

a deviančí. Hned na počátek poznamenejme, že v případě modelů četností pojistných událostí nejsou rezidua tak užitečným nástrojem jako v klasické lineární regresi, neboť, jak je poznamenáno v úvodu článku [14], obecně nemají asymptoticky při $n \to \infty$ normální rozdělení. Všechna výše uvedená rezidua ovšem získávají přibližně normální rozdělení, pracujeme-li s modelem seskupených dat. Upozorněme, že by se v každé z tarifních tříd měl vyskytovat dostatečně vysoký počet smluv (blíže viz článek [14]).

Devianční rezidua jsou v případě poissonovské regrese definovány následovně:

$$r_i^D = \sqrt{2}\text{sign}\,(N_i - \hat{\lambda}_i)\sqrt{N_i \ln \frac{N_i}{\hat{\lambda}_i} - (N_i - \hat{\lambda}_i)}.$$

Zdroj [9] doporučuje pracovat primárně se standardizovanými deviančními rezidui

$$r_i^{Ds} = \frac{r_i^D}{SD_i},$$

kde $SD_i = \sqrt{1 - h_{ii}}$ a h_{ii} je i-tý diagonální prvek matice $\hat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{\frac{1}{2}}$. Hodnoty h_{ii} (leverage) se běžně používají k identifikaci odlehlých pozorování. Pozorování i považujeme za odlehlé, je-li splněno kritérium $h_{ii} > 2p/(n-2p)$. Při kategorizaci proměnných odpadá obvykle nutnost ověřovat toto kritérium. Sečtením všech $(r_i^D)^2$ dostáváme devianci $D(\boldsymbol{k}, \hat{\boldsymbol{\lambda}})$, jde tedy o jakési příspěvky smluv k $D(\boldsymbol{k}, \hat{\boldsymbol{\lambda}})$. Po seskupení dat mají standardizovaná devianční rezidua za platnosti modelu poissonovské regrese přibližně standardizované normální rozdělení. Pro grafické znázornění pak lze využít například histogramu a k verifikaci normality některého ze známých testů (např. Shapirův–Wilkův, Kolmogorův–Smirnovův či Jarqueův–Berův). Kniha [13] (kapitola 2.4) uvádí, že lze alternativně pracovat s Anscombovými rezidui, jelikož nabývají téměř identických hodnot jako rezidua devianční.

K trochu jiným účelům slouží Pearsonova rezidua, definovaná pro poissonovskou regresi vztahem:

$$r_i^p = \frac{N_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}.$$

Pokud skutečně pro odezvy N_i zobecněného lineárního modelu platí $N_i \sim Po(\lambda_i)$, pak $\text{Var}\, N_i = \lambda_i$ a Pearsonova χ^2 statistika

$$\mathcal{X}^2 = \sum_{i=1}^{n} (r_i^p)^2$$

po vydělení $n - p$ konverguje v pravděpodobnosti při $n \to \infty$ k hodnotě disperzního parametru φ, tj. k jedné. Tato statistika je důležitá především v případech, kdy není splněn některý ze základních předpokladů poissonovského modelu, jak bude patrné z poznatků uvedených v příští kapitole. Vydělením r_i^p pomocí výše zmíněného výrazu $SD_i = \sqrt{1 - h_{ii}}$ dostáváme standardizovanou formu Pearsonových reziduí mající dle [9] přibližně jednotkový rozptyl. Lze je také interpretovat jako vhodně normované rozdíly počtů škodných událostí a odhadu jejich střední hodnoty. Jelikož platí

$$\text{corr}\,(N_i, N_j) = \frac{\text{E}\,(N_i - \lambda_i)(N_j - \lambda_j)}{\sqrt{\varphi\lambda_i}\sqrt{\varphi\lambda_j}} \doteq \frac{1}{\varphi}\text{E}\, r_i^p r_j^p,$$

považují za nejvhodnější pro odhad korelací mezi závisle proměnnými v rámci GEE metody, o které bude ještě blíže pojednáno.

2.2.4 Informační kritéria

Při porovnávání více modelů je rozumné podobně jako v případě klasické lineární regrese zkontrolovat také hodnoty některého z informačních kritérií. Mezi různými modely sestavenými na základě stejných dat indikují nižší hodnoty těchto kritérií vyšší kvalitu. Uveďme si nyní několik konkrétních příkladů:

1. Akaikeho informační kritérium AIC:

$$AIC = -2L(\hat{\boldsymbol{\beta}}) + 2p.$$

2. Akaikeho informační kritérium $AICC$:

$$AICC = -2L(\hat{\boldsymbol{\beta}}) + 2p\frac{n}{n-p-1}.$$

3. kritérium BIC:

$$BIC = -2L(\hat{\boldsymbol{\beta}}) + p\ln(n).$$

V praxi se obvykle (dle [15] sekce 4.19, str. 63) preferují kritéria AIC a $AICC$, které mírněji penalizují vysoký počet parametrů, než BIC. Rozdíl mezi AIC a $AICC$ je při vysokém počtu pozorování n pouze nepatrný.

Kapitola 3

Modifikace základního modelu poissonovské regrese

Tato kapitola byla sepsána zejména na základě kapitoly 2 knihy [4].

3.1 Zobecnění modelu poissonovské regrese

Nyní pojednáme o asymptotických vlastnostech maximálně věrohodného odhadu, není-li splněn některý ze základních požadavků. Na počátek poznamenejme, že korektní formulace vět by navíc vyžadovaly splnění tzv. podmínek regularity (viz článek [17]). V této sekci bylo primárně čerpáno z podsekce 3.5.2 knihy [4]. Jednou z možností, jak zobecnit model poissonovské regrese, je nahradit předpoklad o rozdělení (požadavek 3 ve formulaci modelu) pouze podmínkou $\operatorname{Var} N_i = \varphi\lambda_i$ ($\lambda_i = \operatorname{E} N_i$) pro rozptyl, kde φ je disperzní parametr. Je-li navíc $\varphi > 1$, pak mluvíme o modelu nadprůměrné disperze. V literatuře se pro ně často užívá označení ODP (overdisperzion). Odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ probíhá naprosto identicky, neboť zůstává dle [5] (podsekce 3.5.2) stále konzistentní a asymptoticky eficientní. Při dostatečně vysokém počtu pozorování n má $\hat{\boldsymbol{\beta}}$ přibližně normální rozdělení, tj., platí:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx N(0, \varphi\mathcal{I}_n^{-1}), \tag{3.1.1}$$

kde $\mathcal{I}_n = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^T$. Při aplikacích se dle [13] (podsekce 9.2.3, str. 328) φ obvykle odhaduje Pearsonovou χ^2 statistikou vydělenou $n - p$, tj. výrazem $\mathcal{X}^2/(n-p)$.

Další možností, jak zobecnit oba výše zmíněné modely, je plně vypustit požadavek 3 ve formulaci modelu, ale samozřejmě předpis pro střední hodnotu musí být nutně vždy platný. I v takovém případě zůstává totiž $\hat{\boldsymbol{\beta}}$ stále konzistentní. To vysvětluje, proč je poissonovská regrese tak užitečná: Maximálně věrohodný odhad poskytuje při dostatečném počtu pozorování rozumné hodnoty bez ohledu na to, zda skutečné rozdělení je či není Poissonovo a dokonce ani nemusí být exponenciálního typu (viz podsekce 3.5.2 knihy [4]).

Není-li splněn předpoklad o rozdělení, nelze užívat testy poměrem věrohodností, ale je možno aplikovat Waldovy testy, nahradíme-li inverzi Fisherovy informační

matice robustním odhadem $\Sigma_{\hat{\boldsymbol{\beta}}}$, který v sobě implicitně obsahuje empirický odhad rozptylu.[1] Pro dostatečně velké n je varianční matice $\hat{\boldsymbol{\beta}}$ přibližně rovna

$$\Sigma_{\hat{\boldsymbol{\beta}}} = \mathcal{I}_n^{-1}\mathcal{J}\mathcal{I}_n^{-1},$$

kde

$$\mathcal{I}_n = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T d_i \exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) \text{ a } \mathcal{J} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T \operatorname{Var} N_i.$$

V praxi se pak přirozeně $\Sigma_{\hat{\boldsymbol{\beta}}}$ nahrazuje odhadem

$$\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} = \hat{\mathcal{I}}_n^{-1}\hat{\mathcal{J}}\hat{\mathcal{I}}_n^{-1}, \tag{3.1.2}$$

kde

$$\hat{\mathcal{I}}_n = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T \hat{\lambda}_i \text{ a } \hat{\mathcal{J}} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T (\hat{\lambda}_i - k_i)^2,$$

přičemž $\hat{\lambda}_i = d_i \exp\left(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)$. Poznamenejme, že v ODP modelu, kde $\operatorname{var} N_i$ je rovna $\varphi\lambda_i$, přejde robustní odhad varianční matice $\hat{\boldsymbol{\beta}}$ v $\varphi\mathcal{I}_n^{-1}$, neboť

$$\Sigma_{\hat{\boldsymbol{\beta}}} = \mathcal{I}_n^{-1}\mathcal{J}\mathcal{I}_n^{-1} = \mathcal{I}_n^{-1}\varphi\mathcal{I}_n\mathcal{I}_n^{-1} = \varphi\mathcal{I}_n^{-1}.$$

V obou případech má $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ stále při dostatečném počtu pozorování přibližně normální rozdělení s nulovou střední hodnotou a rozptylem $\Sigma_{\hat{\boldsymbol{\beta}}}$ (viz podsekce 3.5.2 knihy [4]).

3.2 Nadprůměrná disperze (overdispersion)

V většině případů odhad disperzního parametru φ převyšuje výrazněji hodnotu jedna, což indikuje, že rozptyl nezávisle proměnných překračuje jejich střední hodnotu. Tomuto jevu říkáme nadprůměrná disperze (overdispersion). Často se také projevuje tím, že rozptyl součtu počtu pojistných událostí v rámci tarifní třídy převyšuje střední hodnotu. Nabízí se dvě možnosti jeho zdůvodnění. Tou první je nedostatečná homogenita tarifních tříd a druhou pozitivní korelace mezi nezávisle proměnnými. Zohledněním těchto vlivů při odhadech se budeme zabývat po zbytek této kapitoly. Oba vyvolávají umělé snížení velikosti odhadů směrodatných odchylek parametrů a zároveň v důsledku také P-hodnot. Poissonovská regrese má tendenci produkovat modely s větším množstvím proměnných, než je ve skutečnosti třeba.

3.2.1 Modelování heterogenity

Pro účely následujícího výkladu budeme považovat regresory za náhodné veličiny, přičemž stále předpokládáme nezávislost odezev N_i. Mnoho nezávisle proměnných je

[1] Termín robustní zde vyjadřuje fakt, že není vyžadováno, aby odezvy měly Poissonovo rozdělení. V literatuře se pro robustní odhad také někdy používá označení sandwichový.

apriorně nepozorovatelných jako například agresivita řidiče či počet ujetých kilometrů za rok a jiné zase nelze zohlednit z důvodů ekonomických, právních či morálních.[2] Nechť $\mathbf{Z}_i$ označuje vektor nezohlednitelných charakteristik příslušejících ke smlouvě i, které ovlivňují škodní frekvenci. Přirozeně složky náhodného vektoru $\mathbf{Z}_i$ mohou být korelovány s regresory. Abychom brali v úvahu také tyto korelace, můžeme složky $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{i\,\dim(\mathbf{Z}_i)})$ pro každé $i = 1, \ldots, n$ vyjádřit pomocí lineární regrese ve tvaru

$$Z_{ij} = \delta_0 + \sum_{k=1}^{p-1} \delta_k X_{ik} + \epsilon_{ij}, \qquad j = 1, \ldots, \dim(\mathbf{Z}_i),$$

Lineární prediktor tak přejde v

$$\begin{aligned} \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \sum_{j=1}^{\dim(\mathbf{Z}_i)} \gamma_j Z_{ij} &= \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \sum_{j=1}^{\dim(\mathbf{Z}_i)} \gamma_j (\delta_0 + \sum_{k=1}^{p-1} \delta_k X_{ik} + \epsilon_{ij}) \\ &= \tilde{\beta}_0 + \sum_{j=1}^{p-1} \tilde{\beta}_j X_{ij} + \tilde{\epsilon}_i \end{aligned}$$

pro vhodné $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \ldots, \tilde{\beta}_{p-1})^T$ a $\tilde{\epsilon}_i$. Odhad $\tilde{\boldsymbol{\beta}}$ tak má v sobě implicitně zahrnut kromě přímého vlivu nezávisle proměnných také jejich korelace se složkami $\mathbf{Z}_i$.

Nyní již můžeme formulovat tzv. smíšený Poissonův model, podle kterého mají N_i Poissonovo rozdělení tvaru

$$(N_i | \mathbf{X}_i = \mathbf{x}_i) \sim Po(d_i \exp(\tilde{\beta}_0 + \sum_{j=1}^{p-1} \tilde{\beta}_j x_{ij} + \tilde{\epsilon}_i)), \qquad i = 1, \ldots, n,$$

kde náhodné veličiny $\tilde{\epsilon}_i$ jsou vzájemně nezávislé a zároveň nezávislé na $\mathbf{X}_i$ pro $i = 1, \ldots, n$, neboť představují reziduální vliv nezohlednitelných charakteristik. Tarifní třídy pak označujeme jako heterogenní, jelikož v nich existují různí řidiči, jejichž škodní frekvence se sice řídí Poissonovým rozdělením, ale s různými středními hodnotami lišícími se v závislosti na $\tilde{\epsilon}_i$.

Odhad $\tilde{\boldsymbol{\beta}}$ se nejčastěji provádí podobně jako v případě zobecněných lineárních modelů také pomocí metody maximální věrohodnosti. V souladu s doposud užívaným značením ($\lambda_i = d_i \exp(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij})$) je střední hodnota počtu pojistných událostí na smlouvě i s apriorně známými hodnotami regresorů $x_{i1}, \ldots, x_{ip-1}$ nyní nově náhodnou veličinou $\lambda_i \Theta_i$, kde $\Theta_i = \exp(\tilde{\epsilon}_i)$ má za cíl modelovat fluktuace okolo střední hodnoty $\lambda_i = d_i \exp(\tilde{\beta}_0 + \sum_{j=1}^{p-1} \tilde{\beta}_j x_{ij})$, a proto požadujeme, aby pro střední hodnotu Θ_i platilo: $\mathrm{E}\,\Theta_i = 1$, pro $i = 1, \ldots, n$. Říkáme pak, že N_i pocházejí ze smíšeného (mixed) Poissonova rozdělení a nadále budeme pro tuto skutečnost používat následující symbolický zápis:

$$N_i \sim \mathcal{M}Po(\lambda_i, \Theta_i).$$

[2] Např. dle směrnic Evropské unie se nesmí ve výsledných sazebnících rozlišovat pohlaví.

Rozptyl N_i pak zřejmě překračuje střední hodnotu, neboť při $\mathrm{E}\,\Theta_i = 1$ platí:

$$\operatorname{var} N_i = \mathrm{E}\,(\underbrace{\operatorname{var}(N_i|\Theta_i)}_{\lambda_i\Theta_i}) + \operatorname{var}(\underbrace{\mathrm{E}\,(N_i|\Theta_i)}_{\lambda_i\Theta_i}) = \lambda_i + \lambda_i^2 \operatorname{var}(\Theta_i) > \lambda_i. \tag{3.2.1}$$

Jako rozdělení náhodné veličiny Θ_i se nejčastěji volí Gama případně log-normální či inverzní Gaussovo (viz závěr sekce 3.4 knihy [5]).

V praxi jsou odhady škodních frekvencí v druhé aposteriorní fázi (viz začátek kapitoly (2)) korigovány pomocí Bayesovských metod v závislosti na postavení řidičů v rámci systému bonus-malus, které by mělo z velké části reflektovat právě vliv $\tilde{\epsilon}_i$ (viz kapitola 4 knihy [4]). Poznamenejme, že dle [4] (předmluva, str. xix) je znalost dostatečně dlouhého historického škodního průběhu klienta pojišťovny cennější z hlediska schopnosti vysvětlit nezávisle proměnnou, než informace o jakékoli apriorní rizikové charakteristice.

3.2.2 Detekce nadprůměrné disperze

Přirozeně pro získání hrubé představy o tom, jakou rozptylovou funkci ve zobecněném lineárním modelu zvolit, je rozumné graficky znázornit empirické odhady rozptylů vůči empirickým středním hodnotách v jednotlivých tarifních třídách. Mějme stejně jako výše (podsekce 2.2.2) tarifní třídy $1, \ldots, q$ a nechť smlouva i spadá do tarifní třídy j pravě tehdy, když platí $score_i = s_j$. Pro tarifní třídu j, označme

$$\hat{m}_j = \frac{\sum_{i|score_i=s_j} k_i}{\sum_{i|score_i=s_j} d_i} \quad \text{a} \quad \hat{\sigma}_j^2 = \frac{\sum_{i|score_i=s_j} (k_i - d_i\hat{m}_j)^2}{\sum_{i|score_i=s_j} d_i}, \quad j = 1, \ldots, q,$$

kde k_i je stejně jako výše počet pojistných událostí na smlouvě i a d_i příslušná expozice v riziku. Podezření na výskyt nadprůměrné disperze vzniká, pokud v podstatné většině případů $\hat{\sigma}_j^2$ převyšuje $\hat{m}_j$. Jestliže dvojice $[\hat{m}_j, \hat{\sigma}_j^2]$ vykazují v grafu přibližně lineární trend se směrnicí 1, pak je pravděpodobně vhodné užít poissonovskou regresi, vyšší velikost směrnice nebo přibližný kvadratický trend indikuje ODP model nebo smíšený Poissonův model.

Samozřejmě grafické znázornění je pouze orientační. Chceme-li skutečně prokázat výskyt nadprůměrné disperze, je třeba zkonstruovat vhodný statistický test. Dle (3.2.1) má rozptylová funkce tvar

$$\operatorname{var}(N_i) = \lambda_i + \tau\lambda_i^2,$$

kde $\tau = \operatorname{var}(\Theta_i)$. Můžeme testovat hypotézu $H_0 : \tau = 0$, která odpovídá poissonovské regresi, proti alternativě $H_1 : \tau > 0$. Testová statistika

$$T = \frac{\sum_{i=1}^n ((k_i - \hat{\lambda}_i)^2 - k_i)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{\lambda}_i^{-2}((k_i - \hat{\lambda}_i)^2 - k_i)^2}\sqrt{\sum_{i=1}^n \hat{\lambda}_i^2}}$$

má za platnosti nulové hypotézy dle [4] (podsekce 2.4.6) asymptoticky standardizované normální rozdělení.

3.3 Negativně binomický regresní model

V předešlé sekci jsme se seznámili s možností, jak se vypořádat s nadprůměrnou disperzí prostřednictvím zavedení nezávislých náhodných veličin $\Theta_1, \ldots, \Theta_n$ modelujících heterogenitu, o kterých budeme dále navíc předpokládat, že jsou stejně rozdělené. Dle této koncepce mají počty škod N_i Poissonovo rozdělení pouze za podmínky, kdy známe kromě rizikových charakteristik také realizaci Θ_i, tj.

$$(N_i|\Theta_i = \theta) \sim Po(\lambda_i\theta), \qquad i = 1, \ldots, n.$$

Do podmínky nyní nebyl zahrnut vektor $\mathbf{X}_i$, neboť $x_{i1}, \ldots, x_{ip-1}$ automaticky považujeme za apriorně známé konstanty. Hodnota θ vystihuje rizikovost smlouvy v rámci tarifní třídy, například $\theta < 1$ indikuje spíše nižší riziko ve srovnání s průměrem. Po zbytek této sekce se budeme zabývat pravděpodobně nejběžněji se v literatuře vyskytujícím případem, kdy $\Theta_1, \ldots, \Theta_n$ mají gama rozdělení. Ukážeme že, ve výsledku tak získáme negativně binomický regresní model.

Nechť hustota náhodné veličiny Θ_i je tvaru:

$$f_{\Theta_i}(\theta) = \frac{a^{\tilde{p}}}{\Gamma(\tilde{p})}\theta^{a-1}\exp(-a\theta), \quad \theta_i \in \mathbb{R}^+,$$

kde a a $\tilde{p}$ jsou kladné parametry, pak $\mathrm{E}\,\Theta_i = \tilde{p}/a$ a $\mathrm{var}\,\Theta_i = \tilde{p}/a^2$ (viz (1.1.1)). Pro splnění požadavku $\mathrm{E}\,\Theta_i = 1$, je nutné, aby si parametry a a $\tilde{p}$ byly rovny. V této chvíli již můžeme odvodit pravděpodobnostní rozdělení N_i. Postupnými úpravami dostáváme

$$\begin{aligned}
P[N_i = k_i] &= \int_0^\infty P[N_i = k_i|\Theta_i = \theta] f_{\Theta_i}(\theta)\mathrm{d}\theta \\
&= \int_0^\infty e^{-\theta\lambda_i}\frac{(\theta\lambda_i)^{k_i}}{k_i!}\frac{a^a}{\Gamma(a)}\theta^{a-1}\exp(-a\theta)\mathrm{d}\theta \\
&= \frac{a^a\lambda_i^{k_i}}{\Gamma(a)k_i!}\int_0^\infty \theta^{k_i+a-1}e^{-(\lambda_i+a)\theta}\mathrm{d}\theta \\
&= \frac{a^a\lambda_i^{k_i}}{\Gamma(a)k_i!}\frac{\Gamma(k_i+a)}{(\lambda_i+a)^{k_i+a}} \\
&= \binom{a+k_i-1}{k_i}\left(\frac{\lambda_i}{a+\lambda_i}\right)^{k_i}\left(\frac{a}{a+\lambda_i}\right)^{a}.
\end{aligned}$$

Z posledního vyjádření je již patrné, že N_i mají negativně binomické rozdělení $NB(a, \lambda_i)$ definované v podsekci 1.1.1.

Negativně binomické rozdělení není obecně rozdělením exponenciálního typu, ale stává se jím, známe-li hodnotu parametru a. Teprve pak je možné mluvit o zobecněném lineárním modelu. Poznamenejme, že tvar rozptylové funkce byl odvozený v podsekci 1.1.1 a plně koresponduje s výsledkem pro smíšené Poissonovo rozdělení při $\mathrm{var}\,\Theta_i = 1/a$. Právě převrácenou hodnotu $1/a$, v literatuře nejčastěji označovanou κ, nazýváme disperzní parametr negativně binomického rozdělení a jeho odhad bývá obvykle uveden ve výstupech při použití některého ze softwarových balíků. Na

skutečné hodnotě disperzního parametru φ rozptyl nijak nezávisí, neboť ta je vždy rovna jedné, proto κ jakýmsi způsobem nahrazuje funkci φ.

Odhad $\hat{\kappa}$ parametru κ lze získat pomocí momentové metody, neboť platí:

$$\operatorname{var} N_i = \lambda_i + \kappa\lambda_i^2 \Rightarrow \kappa = \frac{\operatorname{var} N_i - \lambda_i}{\lambda_i^2},$$
$$\hat{\kappa} = \frac{1}{\hat{a}} = \frac{\sum_{i=1}^n ((k_i - d_i \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}))^2 - k_i)}{\sum_{i=1}^n (d_i \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}))^2},$$

kde $\hat{\boldsymbol{\beta}}$ je konzistentní odhad $\boldsymbol{\beta}$ spočtený na základě poissonovské regrese. Takovýto odhad může být nepřesný a v praxi se spíše používá pouze pro určení počáteční hodnoty iteračního algoritmu, který řeší soustavu věrohodnostních rovnic

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}} L(\boldsymbol{\beta}, a) &= 0, \\ \frac{\partial}{\partial a} L(\boldsymbol{\beta}, a) &= 0. \end{aligned}$$

Poté, co získáme kořen $\tilde{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}, \hat{a})$ této soustavy, můžeme testovat hypotézy o parametrech $\beta_0, \ldots, \beta_{p-1}, a$ pomocí Waldova testu, jak je ilustrováno v knize [4] (podsekce 2.5.2), tj. (2.2.2) přejde v

$$W = (\mathbf{C}\tilde{\boldsymbol{\beta}})^T (\mathbf{C}\tilde{\mathcal{I}}_n^{-1}\mathbf{C}^T)^{-1} (\mathbf{C}\tilde{\boldsymbol{\beta}}) \approx \chi_k^2,$$

kde $\tilde{\mathcal{I}}_n$ je odhadem příslušné Fisherovy informační matice. Alternativně lze využít deviančního testu (blíže viz [4] podsekce 2.4.6).

Jelikož negativně binomické rozdělení přechází při $a \to \infty$ (resp. $\kappa \to 0+$) v Poissonovo, lze podle [15] (sekce 6.2, str. 91) sestavit následující statistický test:

$$\begin{aligned} &H_0 : \kappa = 0, \text{ platí Poissonův model,} \\ &H_1 : \kappa > 0, \text{ platí negativně binomický model,} \end{aligned}$$

kde diperzní parametr κ je tentokrát považován za nezáporný. Platnost H_1 můžeme ověřit na základě modifikované formy deviančního testu. Označme L_{NB} a L_P hodnoty příslušných globálních maxim logaritmických věrohodnostních funkcí v negativně binomickém a Poissonově modelu. Testová statistika

$$2[L_{NB} - L_P] \tag{3.3.1}$$

za platnosti H_0 nabývá hodnoty 0 s pravděpodobností 1/2 a se zbylou pravděpodobností odpovídá hodnotě náhodné veličiny s χ_1^2 rozdělením. Hypotézu H_0 zamítáme na hladině α, pokud testová statistika překročí $\chi_{1,1-2\alpha}^2$, tj. $(1-2\alpha)$ kvantil rozdělení χ_1^2.

Poukažme ještě na jeden z interpretačních rozdílů vůči Poissonově modelu. Věrohodnostní funkce má tvar:

$$\begin{aligned} \ell(\boldsymbol{\beta}, a) &= \prod_{i=1}^n \binom{a + k_i - 1}{k_i} \left(\frac{\lambda_i}{a+\lambda_i}\right)^{k_i} \left(\frac{a}{a+\lambda_i}\right)^a \\ &= \prod_{i=1}^n \frac{\lambda_i^{k_i}}{k_i!} \left(\frac{a}{a+\lambda_i}\right)^a (a+\lambda_i)^{-k_i} \frac{\Gamma(a+k_i)}{\Gamma(a)}, \end{aligned}$$

kde $\lambda_i = d_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})$. Logaritmováním dostáváme:

$$L(\boldsymbol{\beta}, a) = \sum_{i=1}^{n} \log \left(\frac{\lambda_i^{k_i}}{k_i!} \left(\frac{a}{a+\lambda_i} \right)^a (a+\lambda_i)^{-k_i} \frac{\Gamma(a+k_i)}{\Gamma(a)} \right).$$

Soustava rovnic $\frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}, a)$, pak má vyjádření

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}, a) &= \sum_{i=1}^{n} \left(k_i \frac{\partial}{\partial \boldsymbol{\beta}} (\log \lambda_i) - a \frac{\partial}{\partial \boldsymbol{\beta}} (\log (a+\lambda_i)) - k_i \frac{\partial}{\partial \boldsymbol{\beta}} (\log (a+\lambda_i)) \right) \\ &= \sum_{i=1}^{n} \left(k_i \mathbf{x}_i - (a+k_i) \frac{\partial}{\partial \boldsymbol{\beta}} \log (a+\lambda_i) \right) = \sum_{i=1}^{n} \mathbf{x}_i \left(k_i - \lambda_i \frac{a+k_i}{a+\lambda_i} \right). \end{aligned}$$

Oproti věrohodnostním rovnicím pro případ Poissonova rozdělení (viz 2.1.3) došlo k nahrazení λ_i výrazem $\lambda_i \frac{a+k_i}{a+\lambda_i}$. Čtenář dobře obeznámený s Bühlmannovým modelem (viz např. [4] kapitola 3) by si mohl povšimnout, že $\mathrm{E}\left[\Theta_i | N_i = k_i\right] = \frac{a+k_i}{a+\lambda_i}$. Hustota náhodného vektoru (N_i, Θ_i) je rovna

$$\begin{gathered} f_{(N_i,\Theta_i)}(k_i, \theta) = e^{-\theta\lambda_i} \frac{(\theta\lambda_i)^{k_i}}{k_i!} \frac{a^a}{\Gamma(a)} \theta^{a-1} \exp(-a\theta) = \\ konst. \cdot \theta^{k_i+a-1} e^{-(\lambda_i+a)\theta}. \end{gathered}$$

Při daném $N_i = k_i$ přejde hustota Θ_i dle Bayesovy věty v

$$\begin{gathered} f_{(\Theta_i|N_i=k_i)}(\theta|k_i) = \frac{\theta^{k_i+a-1} e^{-(\lambda_i+a)\theta}}{\int_0^\infty \xi^{k_i+a-1} e^{-(\lambda_i+a)\xi} \mathrm{d}\xi} = \\ \frac{(a+\lambda_i)^{a+k_i}}{\Gamma(a+k_i)} \theta^{k_i+a-1} e^{-(\lambda_i+a)\theta}, \end{gathered}$$

což není nic jiného než hustota rozdělení $\Gamma(a+k_i, a+\lambda_i)$, které má střední hodnotu rovnu právě

$$\frac{a+k_i}{a+\lambda_i}.$$

Oproti věrohodnostním rovnicím v Poissonově modelu tak dojde k obohacení střední hodnoty λ_i o informaci obsaženou v počtu škod k_i.

3.4 Panelová data

Někdy je rozumné rozšířit datový vzorek o další sledovaná období. To může být velmi užitečné, pokud došlo v daném roce k nějakému mírnému výkyvu v odhadu $\boldsymbol{\beta}$, způsobeném například dlouhým zimním obdobím. Doposud uvedené modely vycházely z předpokladu nezávislosti, který v tomto případě již není reálný vzhledem k výskytu stejných smluv v různých časových obdobích, a proto je potřeba doposud představené přístupy patřičně upravit.

Nadále předpokládejme, že portfolio tvoří n smluv, nyní ovšem sledovaných během T_i časových období. Nechť N_{it} označuje počet pojistných událostí vzniklých na smlouvě i v období t, kde $i = 1, \ldots, n$, $t = 1, \ldots, T_i$, a d_{it} jsou příslušné expozice v riziku. Data tak získávají panelovou strukturu. Nezávislost lze nyní předpokládat mezi celými náhodnými vektory $\mathbf{N}_1 = (N_{11}, \ldots, N_{1T_1})^T, \ldots, \mathbf{N}_n = (N_{n1}, \ldots, N_{nT_n})^T$, ale ne mezi jejich složkami. Počet smluv n dosahuje obvykle velikosti desetitisíců až milionů, kdežto $T_i, i = 1, \ldots, n$ se pohybují zpravidla v řádu jednotek. Dále se budeme zabývat zobecněním poissonovské regrese v případě, kdy máme k dipozici takováto panelová data.

Před odhadnutím modelu pro panelová data je rozumné vždy zhodnotit výsledky poissonovské regrese rok po roce a zjistit tak, zda dané regresory zůstávají stabilně významné po všechna časová období. Ty nestabilní pak můžeme z modelu vyloučit. V některých případech je patrný ve vývoji odhadu koeficientů jakýsi trend, poté lze často vylepšit kvalitu modelu zahrnutím proměnné zohledňující čas.

3.4.1 Detekce závislostí

O tom, že opravdu existuje mezi složkami jednotlivých náhodných vektorů $\mathbf{N}_1, \ldots, \mathbf{N}_n$ jistá závislost, se lze orientačně přesvědčit zařazením tzv. zpožděné hodnoty vysvětlované proměnné mezi regresory. Konkrétně bychom mohli například vytvořit model poissonovské regrese, v němž by N_{it} byla vysvětlována pomocí vektoru $\mathbf{x}_{it}$ obsahujícím informaci o rizikových charakteristikách a navíc zpožděné hodnoty $N_{i\,t-1}$. Identifikaci proměnné $N_{i\,t-1}$ jako významné můžeme považovat za prokázaní závislosti mezi odezvami.

Výše uvedený postup lze také mírně modifikovat. Nechť $\hat{\boldsymbol{\beta}}$ je odhad vektoru regresních koeficientů $\boldsymbol{\beta}$ pomocí poissonovské regrese, kde platí

$$\mathrm{E}\, N_{it} = d_{it}\exp(\mathbf{x}_{it}^T\boldsymbol{\beta}), \quad i = 1, \ldots, n, t = 1, \ldots, T_i.$$

Odhad $\hat{\boldsymbol{\beta}}$ zůstává konzistentní, i když jsou nezávislé pouze náhodné vektory $\mathbf{N}_1, \ldots, \mathbf{N}_n$, ale ne jejich složky. Proto již stačí pro identifikaci toho, zda je zpožděná hodnota $N_{i\,t-1}$ významná, využít zjednodušeného modelu s předpisem

$$\mathrm{E}\, N_{it} = d_{it}\exp(\mathbf{x}_{it}^T\hat{\boldsymbol{\beta}})\exp(\tilde{\beta}_0 + \tilde{\beta}_1 N_{i\,t-1}),$$

kde offset tentokráte tvoří součet $\ln d_{it} + \mathbf{x}_{it}^T\hat{\boldsymbol{\beta}}$ a odhadovanými parametry jsou $\tilde{\beta}_0$ a $\tilde{\beta}_1$.

3.4.2 GEE metoda odhadu parametrů

Zobecněné rovnice odhadu neboli GEE (general estimating equations) jsou vhodným nástrojem právě v případě, kdy máme k dispozici panelová data. Metoda GEE vychází z velké části z poznatků platných pro tzv. mnohorozměrné zobecněné lineární modely. Veškeré poznatky uvedené v kapitole 2 lze totiž rozšířit i pro případ mnohorozměrného rozdělení exponenciálního typu. Nezávisle proměnné se tak stanou náhodnými vektory, jejichž složky nemusí být nezávislé. V případě zájmu doporučuji

čtenáři nahlédnout do dizertační práce [8], kde jsou základy teorie podrobně popsány včetně důkazů. V této sekci pouze poukážeme na souvislosti, podobnosti a odlišnosti vůči klasickému jednorozměrnému zobecněnému lineárnímu modelu (pro korektnější popis viz [10]).

Vyjděme ze základního modelu poissonovské regrese s nezávislými vysvětlovanými proměnnými. Věrohodnostní rovnice pak mají následující tvar (viz (2.1.3)):

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}(k_{it}-\mathrm{E}\,N_{it})\mathbf{x}_{it}=\sum_{i=1}^{n}\mathbf{X}_i^T(\mathbf{n}_i-\mathrm{E}\,\mathbf{N}_i)=0, \tag{3.4.1}$$

kde $\mathbf{X}_i=(\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT_i})^T$ je matice typu $T_i\times p$ obsahující informaci o rizikových charakteristikách a $\mathbf{n}_i=(k_{i1},\ldots,k_{iT_i})^T$ realizace náhodného vektoru $\mathbf{N}_i$. Varianční matice $\mathbf{N}_i$ je díky nezávislosti diagonální, označme ji

$$\mathbf{A}_i=\begin{pmatrix}\lambda_{i1}&0&\cdots&0\\0&\lambda_{i2}&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\lambda_{iT_i}\end{pmatrix}.$$

Jelikož derivováním dostáváme

$$\frac{\partial}{\partial\boldsymbol{\beta}^T}\mathrm{E}\,\mathbf{N}_i=\begin{pmatrix}\frac{\partial}{\partial\boldsymbol{\beta}^T}\lambda_{i1}\\\vdots\\\frac{\partial}{\partial\boldsymbol{\beta}^T}\lambda_{iT_i}\end{pmatrix}=\begin{pmatrix}\lambda_{i1}\mathbf{x}_{i1}^T\\\vdots\\\lambda_{iT_i}\mathbf{x}_{iT_i}^T\end{pmatrix}=\mathbf{A}_i\mathbf{X}_i,$$

můžeme matici $\mathbf{A}_i$ zakomponovat do vyjádření (3.4.1). Tím přejdeme k zápisu

$$\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}^T}\mathrm{E}\,\mathbf{N}_i\right)^T\mathbf{A}_i^{-1}(\mathbf{n}_i-\mathrm{E}\,\mathbf{N}_i)=0.$$

Nahrazením $\mathbf{A}_i$ skutečnou varianční maticí $\mathbf{N}_i$ (v následujícím výrazu je označena $\mathbf{V}_i$) získáme soustavu věrohodnostních rovnic

$$\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}^T}\mathrm{E}\,\mathbf{N}_i\right)^T\mathbf{V}_i^{-1}(\mathbf{n}_i-\mathrm{E}\,\mathbf{N}_i)=0$$

v mnohorozměrném zobecněném lineárním modelu. Tento výsledek odpovídá zobecnění (1.3.4) a obdobně vyplývá z řetízkového pravidla (pro důkaz viz [8] věta 3.7). Problémem ale stále zůstává, že $\mathbf{V}_i$ není apriorně známa a je ji nutné nahradit vhodným odhadem. Předpokládejme nyní, že platí

$$\mathbf{V}_i=\varphi\mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}, \tag{3.4.2}$$

kde $\mathbf{R}_i(\boldsymbol{\alpha})$ se nazývá pracovní (working) korelační matice a její tvar závisí pouze na parametru $\boldsymbol{\alpha}$, φ je disperzní parametr, jehož účelem je zohlednit případnou

nadprůměrnou disperzi. Uživatel musí ještě před zahájení metody GEE specifikovat strukturu $\mathbf{R}_i(\boldsymbol{\alpha})$. Je-li $\mathbf{R}_i(\boldsymbol{\alpha})$ skutečnou korelační maticí $\mathbf{N}_i$, pak platí také $\operatorname{var}\mathbf{N}_i = \mathbf{V}_i$. Za předpokladu nezávislosti (tj. $\mathbf{R}_i(\boldsymbol{\alpha}) = I_{Ti}$) se vztah (3.4.2) zjednoduší na $\mathbf{V}_i = \varphi\mathbf{A}_i$, což odpovídá při $\varphi > 1$ modelu ODP.

Základem metody GEE je odhadnout regresní koeficienty pomocí řešení soustavy rovnic

$$\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}^T}\boldsymbol{\lambda}_i\right)^T\mathbf{V}_i^{-1}(\mathbf{n}_i-\boldsymbol{\lambda}_i)=0, \tag{3.4.3}$$

kde $\boldsymbol{\lambda}_i = \mathrm{E}\,\mathbf{N}_i$. Bez ohledu na volbu $\mathbf{R}_i(\boldsymbol{\alpha})$ zůstává odhad $\boldsymbol{\beta}$ jakožto kořen kořen rovnice (3.4.3) vždy konzistentní, ale přesto může v praxi dojít vhodným výběrem pracovní korelační matice k výraznému zpřesnění. Uveďme si nyní podrobněji, jaký má metoda GEE průběh (převzato z [16] str. 1987):

Metoda GEE

1. Spočteme odhad $\boldsymbol{\beta}^{(1)}$ parametru $\boldsymbol{\beta}$ pomocí poissonovské regrese za předpokladu nezávislosti. Poté započne iterační proces. Hodnotu k označující pořadí iterace nastavíme na 0.

2. Navýší se hodnota k o jedna a na základě Pearsonových reziduí se stanový odhad $\boldsymbol{\alpha}^{(k)}$ parametru $\boldsymbol{\alpha}$ určující pracovní korelační matici $\mathbf{R}_i(\boldsymbol{\alpha})$ (viz poznámka níže a příští podsekce).

3. Dopočítá se odhad varianční matice $\mathbf{V}_i = \varphi\mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$ pomocí vztahu

$$\mathbf{V}_i(\boldsymbol{\lambda}_i^{(k)}) = \mathbf{A}_i^{1/2}(\boldsymbol{\lambda}_i^{(k)})(\varphi^{(k)}\mathbf{R}_i(\boldsymbol{\alpha}^{(k)}))\mathbf{A}_i^{1/2}(\boldsymbol{\lambda}_i^{(k)}),$$

   kde $\boldsymbol{\lambda}_i^{(k)} = (d_{i1}\exp(\mathbf{x}_{i1}^T\boldsymbol{\beta}^{(k)}),\ldots,d_{iT_i}\exp(\mathbf{x}_{iT_i}^T\boldsymbol{\beta}^{(k)}))^T$, $\mathbf{A}_i(\boldsymbol{\lambda}_i^{(k)})$ je maticovou funkcí odpovídající $diag(\lambda_{i1}^{(k)},\ldots,\lambda_{iT_i}^{(k)})$, a $\varphi^{(k)}$ je odhad disperzního parametru na základě Pearsonovy χ^2 statistiky vycházející z $\boldsymbol{\beta}^{(k)}$ (viz poznámka níže).

4. Aktualizuje se vektor regresních koeficientů pomocí vztahu

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left[\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}^T}\boldsymbol{\lambda}_i^{(k)}\right)^T\mathbf{V}_i^{-1}(\boldsymbol{\lambda}_i^{(k)})\left(\frac{\partial}{\partial\boldsymbol{\beta}^T}\boldsymbol{\lambda}_i^{(k)}\right)\right]^{-1}\mathbb{U}(\boldsymbol{\beta}^{(k)}),$$

   kde

$$\begin{aligned}\mathbb{U}(\boldsymbol{\beta}^{(k)}) &= \left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}^T}\boldsymbol{\lambda}_i^{(k)}\right)^T\mathbf{V}_i^{-1}(\boldsymbol{\lambda}_i^{(k)})(\mathbf{n}_i-\boldsymbol{\lambda}_i^{(k)})\right),\\ \mathbf{V}_i(\boldsymbol{\lambda}_i^{(k)}) &= \mathbf{A}_i^{1/2}(\boldsymbol{\lambda}_i^{(k)})(\varphi^{(k)}\mathbf{R}_i(\boldsymbol{\alpha}^{(k)}))\mathbf{A}_i^{1/2}(\hat{\boldsymbol{\lambda}}_i^{(k)}).\end{aligned}$$

5. Pokud není dosaženo kritéria konvergence $\parallel\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}^{(k-1)}\parallel<\varepsilon$, opakujeme znovu body 2 až 5.

Poznámky k metodě GEE

Ad 2) Jelikož Pearsonova rezidua

$$r_{it}^p = \frac{N_{it} - \hat{\lambda}_{it}}{\sqrt{\hat{\lambda}_{it}}}, \quad i = 1, \ldots, n, t = 1, \ldots, T_i$$

mají přibližně nulovou střední hodnotu a rozptyl φ, platí také pro korelace přibližně

$$\operatorname{corr}(N_{it}, N_{is}) \doteq \operatorname{corr}(r_{it}^p, r_{is}^p) \doteq \frac{1}{\hat{\varphi}} \operatorname{cov}(r_{it}^p, r_{is}^p) \doteq \frac{1}{\hat{\varphi}} \mathrm{E}\, r_{it}^p r_{is}^p,$$

kde $\hat{\varphi}$ je odhad disperzního parametru pomocí Pearsonovy χ^2 statistiky, tj.

$$\hat{\varphi} = \frac{1}{(\sum_{i=1}^n T_i) - p} \sum_{i=1}^{n} \sum_{t=1}^{T_i} (r_{it}^p)^2.$$

Ad 3) Matice $\varphi^{(k)} \mathbf{R}_i(\boldsymbol{\alpha}^{(k)})$ jsou typu $T_i \times T_i$ a obsahují odhad $\mathrm{E}\, r_{it}^p r_{is}^p$ na pozici (t, s), pro $t = 1 \ldots, T_i$ a $s = 1, \ldots, T_i$ (blíže viz následující podsekce). Proto není ve skutečnosti nutné dopočítávat $\varphi^{(k)}$, pokud ovšem není korelační matice $\mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{R}_i$ apriorně pevně zvolena, poté dojde v bodě 2 pouze k navýšení k.

3.4.3 Možné volby struktury pracovní korelační matice

Díky pracovní korelační matici jsou v modelu zohledněny závislosti mezi vysvětlovanými proměnnými. Pokud je $\mathbf{R}_i(\boldsymbol{\alpha})$, jednotkovou maticí řádu T_i, pak metoda GEE odpovídá odhadu v klasické poissonovské regresi zmíněné v kapitole 3 využívající předpokladu nezávislosti. Nyní ukáži některé další možné volby $\mathbf{R}_i(\boldsymbol{\alpha})$ včetně odhadu parametru $\boldsymbol{\alpha}$ a navíc vždy uvedu také anglický termín pro název dané struktury matice (přičemž bylo čerpáno z [16] str. 1985-1986).

1. Pásové korelace řádu m (m-dependance), tj.

$$\operatorname{corr}(N_{it}, N_{i,t+s}) = \begin{cases} 1 & s = 0 \\ \alpha_s & s = 1, \ldots, m \\ 0 & s > m, \end{cases}$$

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \cdots & \alpha_m & 0 & \cdots & 0 \\ \alpha_1 & 1 & \alpha_1 & \ddots & & \ddots & \ddots & \vdots \\ \alpha_2 & \alpha_1 & 1 & & \ddots & & \ddots & 0 \\ \vdots & \ddots & & \ddots & & \ddots & & \alpha_m \\ \alpha_m & & \ddots & & \ddots & & \ddots & \vdots \\ 0 & \ddots & & \ddots & & \ddots & & \alpha_2 \\ \vdots & \ddots & \ddots & & \ddots & & \ddots & \alpha_1 \\ 0 & \cdots & 0 & \alpha_m & \cdots & \alpha_2 & \alpha_1 & 1 \end{pmatrix},$$

pro všechna $i = 1, \ldots, n$. Odhad složek parametru $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$ vypadá následovně:

$$\hat{\alpha}_s = \frac{1}{\hat{\varphi}} \cdot \frac{1}{(\sum_{i=1}^{n} T_i) - ns - p} \sum_{i=1}^{n} \sum_{t=1}^{T_i - s} r_{it}^p r_{i,t+s}^p, \quad s = 1, \ldots, m.$$

2. Stejné korelace (exchangeable), tj.

$$\operatorname{corr}(N_{it}, N_{is}) = \begin{cases} 1 & s = t \\ \alpha_s & s \neq t, \end{cases}$$

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha \\ \alpha & \cdots & \alpha & 1 \end{pmatrix},$$

pro všechna $i = 1, \ldots, n$. Odhad parametru α vypadá následovně:

$$\begin{aligned} \hat{\alpha} &= \frac{1}{\hat{\varphi}} \cdot \frac{1}{T_i^* - p} \sum_{i=1}^{n} \sum_{t<s} r_{it}^p r_{is}^p, \\ T_i^* &= \sum_{i=1}^{n} \frac{T_i(T_i - 1)}{2}. \end{aligned}$$

3. Různé korelace (unstructured), tj.

$$\operatorname{corr}(N_{it}, N_{is}) = \begin{cases} 1 & s = t \\ \alpha_{ts} & s > t \\ \alpha_{st} & s < t, \end{cases}$$

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1T_i} \\ \alpha_{12} & 1 & \alpha_{23} & \ddots & \vdots \\ \alpha_{13} & \alpha_{23} & 1 & & \alpha_{T_i-2,T_i} \\ \vdots & \ddots & & \ddots & \alpha_{T_i-1,T_i} \\ \alpha_{1T_i} & \cdots & \alpha_{T_i-2,T_i} & \alpha_{T_i-1,T_i} & 1 \end{pmatrix},$$

pro všechna $i = 1, \ldots, n$. Odhad složek parametru

$$\boldsymbol{\alpha} = (\alpha_{12}, \ldots, \alpha_{1T_i}, \alpha_{23}, \ldots, \alpha_{2T_i}, \alpha_{34}, \ldots\ldots\ldots, \alpha_{T_i-1,T_i})^T$$

vypadá následovně:

$$\hat{\alpha}_{ts} = \frac{1}{\hat{\varphi}} \cdot \frac{1}{n - p} \sum_{i=1}^{n} r_{it}^p r_{is}^p, \quad t < s, s = 2, \ldots, T_i.$$

4. Autoregrese prvního řádu (AR1), tj.

$$\operatorname{corr}(N_{it}, N_{i\,t+s}) = \alpha^s, \quad s = 0, 1, \ldots, T_i - t,$$

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{T_i-1} \\ \alpha & 1 & \alpha & \ddots & \vdots \\ \alpha^2 & \alpha & 1 & & \alpha^2 \\ \vdots & \ddots & & \ddots & \alpha \\ \alpha^{T_i-1} & \cdots & \alpha^2 & \alpha & 1 \end{pmatrix},$$

pro všechna $i = 1, \ldots, n$. Odhad parametru α vypadá následovně:

$$\begin{aligned} \hat{\alpha} &= \frac{1}{\hat{\varphi}} \cdot \frac{1}{K-p} \sum_{i=1}^{n} \sum_{t<T_i-1} r_{it}^p r_{i,t+1}^p, \\ K &= \sum_{i=1}^{n} (T_i - 1). \end{aligned}$$

Poznámka. Poznamenejme, že v některé statistické programy umožňují také předem specifikovat vlastní fixní pracovní korelační matici nezávislou na parametru $\boldsymbol{\alpha}$.

3.4.4 Testování významnosti parametrů

Hned na počátek poznamenejme, že metoda GEE není založena plně na teorii maximální věrohodnosti, a proto nelze jednoduše aplikovat některý z výše zmíněných testů. Speciálně nelze užívat devianční testy. Pokud je správně specifikován předpis pro střední hodnotu vysvětlované proměnné a zároveň také pracovní korelační matice, pak lze konzistentně odhadnout varianční matici $\hat{\boldsymbol{\beta}}$ výrazem

$$\mathcal{I}_{GEE}^{-1} = \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}^T} \mathrm{E}\, \mathbf{N}_i \right)^T \mathbf{V}_i^{-1} \left(\frac{\partial}{\partial \boldsymbol{\beta}^T} \mathrm{E}\, \mathbf{N}_i \right) \right)^{-1},$$

kde $\mathcal{I}_{GEE}$ odpovídá Fisherově informační matici v mnohorozměrném zobecněném lineárním modelu, neboť platí

$$\mathcal{I}_{GEE} = \operatorname{var} U(\boldsymbol{\beta}) = \operatorname{var} \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}^T} \mathrm{E}\, \mathbf{N}_i \right)^T \mathbf{V}_i^{-1} (\mathbf{N}_i - \mathrm{E}\, \mathbf{N}_i) \right).$$

(i v mnohorozměrném případě platí obdoba (A.1.6), viz věta 1.5 práce [8]). Ovšem varianční matice $\hat{\boldsymbol{\beta}}$ lze odhadnout také robustně pomocí vyjádření

$$\mathcal{I}_{GEE}^{-1} \mathcal{J}_{GEE} \mathcal{I}_{GEE}^{-1},$$

kde

$$\mathcal{J}_{GEE} = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}^T} \mathrm{E}\, \mathbf{N}_i \right)^T \mathbf{V}_i^{-1} \operatorname{var}(\mathbf{N}_i) \mathbf{V}_i^{-1} \left(\frac{\partial}{\partial \boldsymbol{\beta}^T} \mathrm{E}\, \mathbf{N}_i \right).$$

Termín robustní vyjadřuje fakt, že není nutné, aby struktura korelační matice byla správně zvolena, a přesto odhad varianční matice, který bývá běžně označován také jako sandwichový, zůstává konzistentní (dle [4], str. 104). Při výpočtu se pak $\operatorname{var}\mathbf{N}_i$ nahradí

$$\hat{\operatorname{var}}\mathbf{N}_i = (\mathbf{N}_i - \operatorname{E}\hat{\mathbf{N}}_i)(\mathbf{N}_i - \operatorname{E}\hat{\mathbf{N}}_i)^T$$

a zbytek se odhadne na základě GEE metody. Asymptoticky má při splnění relativně slabých předpokladů (blíže viz článek [10]) odhad $\hat{\boldsymbol{\beta}}$ rozdělení $N_p(0, \mathcal{I}_{GEE}^{-1}\mathcal{J}_{GEE}\mathcal{I}_{GEE}^{-1})$, a proto lze užít podobným způsobem jako v podsekci (2.2.1) Waldovy testy. Alternativou je modifikovaná forma skórového testu pro GEE. Ten je založen na faktu, že za platnosti nulové hypotézy $\mathbf{C}\boldsymbol{\beta} = 0$ ($\mathbf{C}$ je matice plné hodnosti typu $k \times p$) má statistika (dle [16] str. 1993)

$$U(\hat{\boldsymbol{\beta}})^T\mathcal{I}_{GEE}^{-1}\mathbf{C}^T(\mathbf{C}\mathcal{I}_{GEE}^{-1}\mathcal{J}_{GEE}\mathcal{I}_{GEE}^{-1}\mathbf{C}^T)^{-1}\mathbf{C}\mathcal{I}_{GEE}^{-1}U(\hat{\boldsymbol{\beta}})$$

přibližně χ_k^2 rozdělení, kde $\hat{\boldsymbol{\beta}}$ je odhad pomocí metody GEE a k počet řádků matice $\boldsymbol{C}$.

Kapitola 4

Numerická studie na reálných datech

Účelem této studie bude aplikovat techniky uvedené v předchozích kapitolách na reálná data prostřednictvím statistického softwaru SAS, který poskytuje uživateli možnost využít procedury GENMOD, jež slouží zejména pro práci se zobecněnými lineárními modely. V příloze B jsou uvedeny některé důležité části zdrojového kódu. Data byla poskytnuta pojišťovnou Kooperativa a týkají se smluv povinného ručení osobních automobilů s počátkem platnosti či datem prolongace (prodloužení platnosti smlouvy) v období let 2008-2011. Kromě počtů pojistných událostí a příslušných expozic v riziku obsahují také hodnoty čtyřech rizikových charakteristik. Těmito charakteristikami jsou:

ts	tarifní skupina (dle objemu motoru),
reg	počet obyvatel v místě bydliště pojistníka,
sv	stáří vozidla,
vek	věk pojištěného.

Jelikož nejsem oprávněn prezentovat způsob kategorizace jednotlivých rizikových charakteristik, zůstanou tyto informace čtenáři zatajeny. Navíc odhady regresních koeficientů a směrodatných odchylek v níže uvedených tabulkách jsou vynásobeny náhodnou konstantou z intervalu $(0, 2)$. Tato úprava totiž neovlivní hodnoty Waldových statistik a příslušných P-hodnot. Poznamenejme ještě, že interpretace výše jednotlivých regresních koeficientů byla již v podsekci 2.1.3 popsána a nyní se zaměříme především na ilustraci verifikačních metod.

V této kapitole budeme pracovat se dvěma datovými soubory. Soubor A obsahuje informace o smlouvách s počátkem platnosti či datem prolongace v roce 2010 a soubor B v letech 2008-2010. Pro odhad modelu poissonovké regrese a negativně binomického modelu využijeme soubor A. Pro aplikaci GEE metody pak budeme vycházet ze souboru B. Jednotlivé expozice smluv budou do modelů zařazeny prostřednictvím offsetu.

Nejprve se na základě odhadů ODP modelů pokusíme identifikovat regresory, které významně ovlivňují závisle proměnnou. Pro tuto skupinu regresorů poté odhadneme a verifikujeme model poissonovské regrese, negativně binomický, ale také i model pro panelová data popsaný v sekci 3.4. Na závěr této kapitoly porovnáme jejich predikční schopnost pro rok 2011.

4.1 Identifikace regresorů

Mějme hodnoty rizikových charakteristik kategorizovány způsobem uvedeným v prvních dvou sloupcích tabulky 4.1. Zbytek tabulky obsahuje informace týkající se odhadu výchozího ODP modelu. Odhady směrodatných odchylek ve čtvrtém sloupci jsou odmocninou součinu odhadu $\hat{\varphi} = \mathcal{X}^2/(n-p)$ disperzního parametru a příslušného diagonálního prvku matice $\hat{\mathcal{I}}_n^{-1} = (\sum_{i=1}^n \hat{\lambda}_i \mathbf{x}_i \mathbf{x}_i^T)^{-1}$(viz (3.1.1)). SAS provede tuto úpravu při nastavení volby `Scale=Pearson` v rámci procedury GENMOD. Hodnoty Waldových testových testových statistik pak odpovídají druhé mocnině podílu odhadu parametru a směrodatné odchylky (viz (2.2.3)). P-hodnota zde udává pravděpodobnost, že realizace náhodné veličiny s rozdělením χ_1^2 bude větší než Waldova testová statistika.

riziková charakteristika	kategorie	odhad regresních koeficientů	odhad směrodatné odchylky	Waldova testová statistika	P-hodnota
intercept		0	0		
ts	1	-3,3256	0,1244	714,31	<0,0001
ts	2	-3,1617	0,1257	632,31	<0,0001
ts	3	-3,0387	0,1258	583,34	<0,0001
ts	4	-2,9131	0,1313	492,36	<0,0001
reg	1	0,2903	0,0182	254,46	<0,0001
reg	2	0,1628	0,0187	75,87	<0,0001
reg	3	0	0		
sv	0	0,0598	0,1475	0,16	0,6852
sv	1	0,7556	0,1254	36,29	<0,0001
sv	2	0,7956	0,1249	40,53	<0,0001
sv	3	0	0		
vek	1	0,5155	0,0495	108,56	<0,0001
vek	2	0,2175	0,0235	85,70	<0,0001
vek	3	0,0421	0,0176	5,72	0,0168
vek	4	0	0		

Tabulka 4.1: Odhad výchozího ODP modelu.

Velikost odhadu disperzního parametru se výrazněji odlišuje od 1, přibližně je rovna $1,5$, což indikuje, že není rozumné používat devianční testy. Pouze jediná z P-hodnot v tabulce 4.1 je větší než 0,05. To můžeme interpretovat tak, že sloučením kategorie 0 rizikové charakteristiky stáří vozidla s referenční kategorií 3 dostaneme podmodel, jehož platnost Waldův test nezamítá. Základní informace o odhadu podmodelu pak dále shrnuje tabulka 4.2, přičemž nová kategorie vzniklá sloučením je zde označena jako 0.

V podmodelu již jsou všechny nezávisle proměnné významné na hladině 0,05. Zlepšení kvality indikuje kritérium AIC, které pokleslo z hodnoty 115443,3 na 115441,6.

riziková charakteristika	kategorie	odhad regresních koeficientů	odhad směrodatné odchylky	Waldova testová statistika	P-hodnota
intercept		0	0		
ts	1	-2,5302	0,0185	18766,70	<0,0001
ts	2	-2,3661	0,0156	22937,22	<0,0001
ts	3	-2,2431	0,0165	18570,75	<0,0001
ts	4	-2,1175	0,0426	2467,99	<0,0001
reg	1	0,2903	0,0182	254,60	<0,0001
reg	2	0,1628	0,0187	75,89	<0,0001
reg	3	0	0		
sv	0	-0,7537	0,0682	122,16	<0,0001
sv	1	-0,0400	0,0162	6,19	0,0135
sv	2	0	0		
vek	1	0,5155	0,0495	108,52	<0,0001
vek	2	0,2176	0,0235	85,71	<0,0001
vek	3	0,0421	0,0176	5,72	0,0168
vek	4	0	0		

Tabulka 4.2: Odhad podmodelu se zohledněním nadprůměrné disperze.

Procedura GENMOD vždy generuje pro uživatele tabulku kritérií pro posouzení kvality modelu, jejíž součástí jsou mimo jiné například informační kritéria AIC, AICC, BIC nebo odhad disperzního parametru $\hat{\varphi} = \mathcal{X}^2/(n-p)$. Rovnost dvojice regresních koeficientů příslušejících jedné rizikové charakteristice lze otestovat jednoduše změnou referenční kategorie. Procedura GENMOD umožňuje nastavit libovolnou referenční kategorii pomocí příkazu `ref` uvedeném v závorce za názvem rizikové charakteristiky. Vyzkoušel jsem nastavení všech možných referenčních skupin rizikových charakteristik, ale P-hodnoty Waldových testů pro jednotlivé parametry byly vždy menší než 0,05. GENMOD také poskytuje možnost využít volby `contrast`, která vyžaduje specifikaci matice hypotézy, a test pak přesně odpovídá popisu v podsekci 2.2.1.

Přesnější představu o výši rozptylu odhadu regresních koeficientů a tedy také P-hodnot Waldových testů lze získat výpočtem robustního odhadu varianční matice (viz (3.1.2)). Robustní odhady se uplatní použitím příkazu `repeated`, který slouží pro aktivaci GEE metody. Pokud navíc nastavíme parametr `type=Ind` (independent), považují se odezvy za nezávislé. Nevýhodou ale je, že v takovém případě může někdy výpočet trvat výrazně delší dobu. Shrnutí údajů souvisejících s robustním odhadem uvádí tabulka 4.3. Testovou statistiku v pátém sloupci uvádí zároveň ve svých výstupech SAS při uvedení příkazu `repeated` a je pouze poměrem odhadu regresního koeficientu a jeho směrodatné odchylky. V příští sekci se o GEE metodě ještě blíže zmíním.

riziková charakteristika	kategorie	odhad regresních koeficientů	odhad směrodatné odchylky	testová statistika	P-hodnota
intercept		0	0		
ts	1	-2,5302	0,0154	-164,65	<0,0001
ts	2	-2,3661	0,0131	-180,75	<0,0001
ts	3	-2,2431	0,0137	-164,00	<0,0001
ts	4	-2,1175	0,0361	-58,67	<0,0001
reg	1	0,2903	0,0151	19,24	<0,0001
reg	2	0,1628	0,0155	10,51	<0,0001
reg	3	0	0		
sv	0	-0,7537	0,0579	-13,02	<0,0001
sv	1	-0,0400	0,0134	-3,00	0,0028
sv	2	0	0		
vek	1	0,5155	0,0408	12,64	<0,0001
vek	2	0,2176	0,0195	11,16	<0,0001
vek	3	0,0421	0,0146	2,89	0,0039
vek	4	0	0		

Tabulka 4.3: Poissonův model s úpravou směrodatných odchylek na základě robustního odhadu.

4.2 Odhad, verifikace a porovnání modelů

4.2.1 Poissonův a negativně binomický model

Pracujme nyní s kategorizací rizikových charakteristik získanou v předešlé sekci (viz např. tabulka 4.3). Shrnutí základních údajů týkajících se odhadu modelu poissonovské regrese je uvedeno v tabulce 4.4. Povšimněme si, že došlo k poklesu odhadů směrodatných odchylek ve srovnání s předchozími modely (viz tabulky 4.3,4.2) a v důsledku také zvýšení Waldových statistik a snížení P-hodnot.

Nejprve se pokusme ověřit, zda model poissonovské regrese odpovídá napozorovaným hodnotám. Seskupíme-li data dle tarifních tříd, pak odhad regresních koeficientů ale i směrodatných odchylek zůstává stále stejný (viz tabulka 4.5). Tato skutečnost vyplývá z poznatků uvedených v závěru sekce 2.2.2. Tvar histogramu (obrázek 4.2.1) standardizovaných deviančních reziduí přibližně připomíná hustotu normálního rozdělení a navíc Shapirův-Wilkův test, jehož P-hodnota je rovna 0,48, nezamítá normalitu standardizovaných reziduí na hladině 0,05.[1] To indikuje vysokou kvalitu modelu.

Na obrázku 4.2.2 jsou znázorněny dvojice $[\hat{m}_j, \hat{\sigma}_j^2]$ (viz 3.2.2). Přibližně v šedesáti procentech případů přesahuje odhad rozptylu $\hat{\sigma}_j^2$ odhad střední hodnoty $\hat{m}_j$. Obrázek napovídá, že by naše data mohla odpovídat modelu poissonovské regrese. Ovšem odhad $\hat{\varphi}$ pomocí Pearsonovy χ^2 statistiky, který má konvergovat v pravděpodobnosti za platnosti modelu k hodnotě 1, je přibližně roven $1,5$. To na druhou stranu indikuje

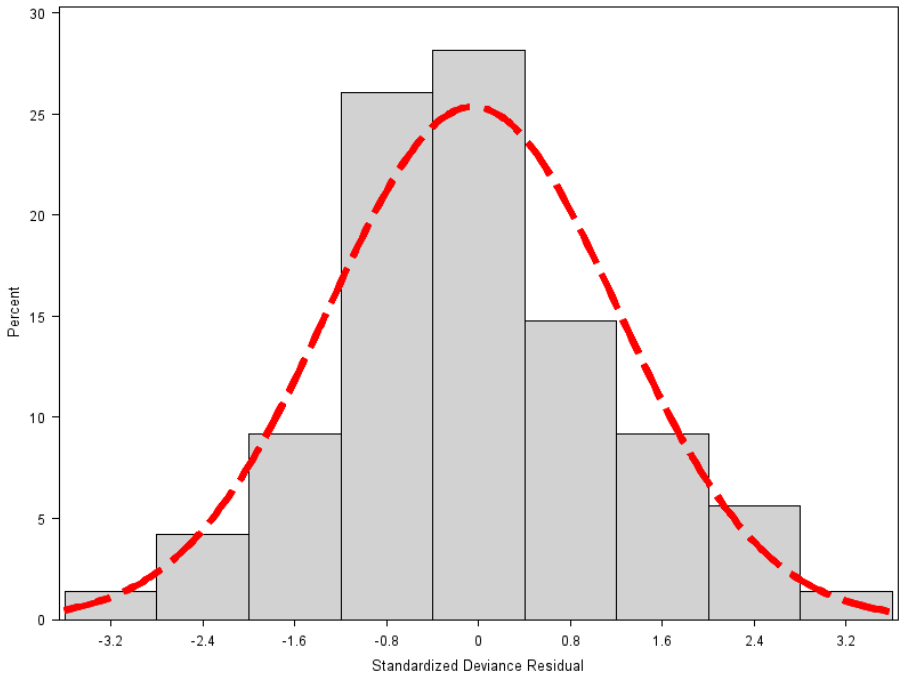
[1] Hodnota testové statistiky Shapirova-Wilkova testu je 0,99.

riziková charakteristika	kategorie	odhad regresních koeficientů	odhad směrodatné odchylky	Waldova testová statistika	P-hodnota
intercept		0	0		
ts	1	-2,5302	0,0150	28327,18	<0,0001
ts	2	-2,3661	0,0127	34622,33	<0,0001
ts	3	-2,2431	0,0134	28031,41	<0,0001
ts	4	-2,1175	0,0347	3725,28	<0,0001
reg	1	0,2903	0,0148	384,31	<0,0001
reg	2	0,1628	0,0152	114,55	<0,0001
reg	3	0	0		
sv	0	-0,7537	0,0555	184,39	<0,0001
sv	1	-0,0400	0,0132	9,22	0,0024
sv	2	0	0		
vek	1	0,5155	0,0402	163,81	<0,0001
vek	2	0,2176	0,0191	129,38	<0,0001
vek	3	0,0421	0,0143	8,64	0,0033
vek	4	0	0		

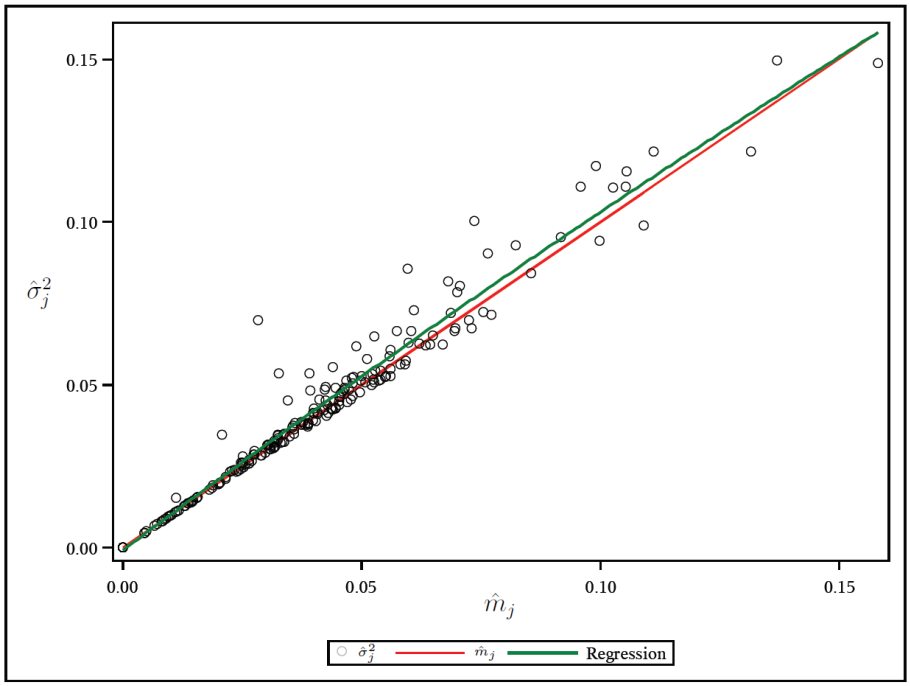
Tabulka 4.4: Odhad modelu poissonovské regrese.

riziková charakteristika	kategorie	odhad regresních koeficientů	odhad směrodatné odchylky	Waldova testová statistika	P-hodnota
intercept		0	0		
ts	1	-2,5302	0,0150	28327,18	<0,0001
ts	2	-2,3661	0,0127	34622,33	<0,0001
ts	3	-2,2431	0,0134	28031,41	<0,0001
ts	4	-2,1175	0,0347	3725,28	<0,0001
reg	1	0,2903	0,0148	384,31	<0,0001
reg	2	0,1628	0,0152	114,55	<0,0001
reg	3	0	0		
sv	0	-0,7537	0,0555	184,39	<0,0001
sv	1	-0,0400	0,0132	9,22	0,0024
sv	2	0	0		
vek	1	0,5155	0,0402	163,81	<0,0001
vek	2	0,2176	0,0191	129,38	<0,0001
vek	3	0,0421	0,0143	8,64	0,0033
vek	4	0	0		

Tabulka 4.5: Odhad modelu poissonovské regrese na základě seskupených dat dle tarifních tříd.



Obrázek 4.2.1: Histogram standardizovaných deviančních reziduí. Červenou čarou je znázorněn tvar hustoty standardizovaného normálního rozdělení.



Obrázek 4.2.2: Porovnání odhadů středních hodnot a rozptylů v tarifních třídách. Červenou barvu má přímka se směrnicí rovnou jedné a zelenou je znázorněn přibližný kvadratický trend.

výskyt nadprůměrné disperze, což zároveň potvrzuje test uvedený v podsekci 3.2.2, který zamítá rovnost střední hodnoty a rozptylu. Jeho testová statistika je rovna 7,14 a P-hodnota přibližně 0. Domnívám se, že velmi nízká P-hodnota je zde důsledkem velké síly testu při řádově statisících pozorování. Tento výsledek napovídá, že by naše data mohla lépe odpovídat negativně binomickému modelu. Poznamenejme ale, že pokud by ve skutečnosti pozorování pocházela z negativně binomického modelu nebo obecněji smíšeného Poissonova modelu, pak by trend znázorněný v obrázku 4.2.2 zelenou barvou měl vykazovat spíše konvexní průběh.

Přejděme nyní k negativně binomickému modelu. Maximálně věrohodný odhad $\hat{\kappa}$ disperzního parametru negativně binomického rozdělení $\kappa = 1/a$ je roven 1,27. Jak je vidět z tabulky 4.6, oproti Poissonově modelu došlo pouze k velmi mírnému zvětšení odhadů směrodatných odchylek, ale odhad regresních koeficientů zůstal přibližně stejný. Opět Waldovy testy zamítají nulovost jednotlivých regresorů na hladině 0,05.

Testová statistika

$$2[L_{NB} - L_P]$$

modifikovaného deviančního testu (viz (3.3.1)) nabývá hodnoty $197,1$, což odpovídá P-hodnotě menší než 0,0001, a proto zamítáme na hladině $0,05$ nulovou hypotézu o platnosti Poissonova modelu ve prospěch negativně binomického. Konkrétní výše logaritmických věrohodnostních funkcí jsou $L_{NB} = -57611,23$ a $L_P = -57709,78$.

riziková charakteristika	kategorie	odhad regresních koeficientů	odhad směrodatné odchylky	Waldova testová statistika	P-hodnota
intercept		0	0		
ts	1	-2,5280	0,0153	27303,05	<0,0001
ts	2	-2,3629	0,0130	33166,75	<0,0001
ts	3	-2,2390	0,0137	26685,90	<0,0001
ts	4	-2,1131	0,0357	3501,87	<0,0001
reg	1	0,2907	0,0152	365,98	<0,0001
reg	2	0,1634	0,0156	110,34	<0,0001
reg	3	0	0		
sv	0	-0,7560	0,0559	182,90	<0,0001
sv	1	-0,0405	0,0135	9,05	0,0026
sv	2	0	0		
vek	1	0,5237	0,0418	156,69	<0,0001
vek	2	0,2187	0,0196	124,00	<0,0001
vek	3	0,0425	0,0146	8,43	0,0037
vek	4	0	0		

Tabulka 4.6: Odhad negativně binomického modelu.

4.2.2 Metoda GEE a porovnání predikčních schopností

Než přejdeme k práci s datovým souborem B, zkusme nejprve detekovat pomocí technik v podsekci 3.4.1, zda vůbec nějaké závislosti mezi odezvami existují. Při za-

řazení zpožděné hodnoty závisle proměnné mezi regresory Waldův test nezamítá její významnost, jelikož příslušná P-hodnota je menší než 0,0001, což můžeme považovat za indikaci existence závislostí mezi odezvami.

Vycházejme nyní z datového souboru B. Pro aplikaci metody GEE pomocí procedury GENMOD je nutné uvést příkaz `repeated` a uživatel má na výběr z několika struktur pracovní korelační matice. Tu specifikujeme pomocí nastavení vhodného parametru volby `type`. Na výběr jsou následující čtyři možnosti:

`type=MDEP(m)`	pásové korelace řádu m (m-dependance),
`type=EXCH`	stejné korelace (exchangeable),
`type=UNSTR`	různé korelace (unstructured),
`type=AR(1)`	autoregrese prvního řádu (AR1).

Jelikož si nyní nejsme jisti, kterou ze struktur využít, předpokládejme různé korelace mezi proměnnými a nastavme `type=UNSTR`. Tabulka 4.7 je částí výstupu získaného užitím GEE metody na datový soubor B (2008-2010) při volbě `type=UNSTR`. Odhady směrodatných odchylek odpovídají odmocninám diagonálních prvků robustní varianční matice $\hat{\boldsymbol{\beta}}$ (viz sekce 3.4.4). Testové statistiky v pátém sloupci jsou poměrem odhadů regresních koeficientů a směrodatných odchylek. P-hodnoty udávají pravděpodobnost, že náhodná veličina s normovaným normálním rozdělení bude v absolutní hodnotě větší než absolutní hodnota testové statistiky.

Nechť $\lambda_j^{pred}, j = 1, \ldots, n^{pred}$ jsou predikce střední hodnoty počtů pojistných událostí pro rok 2011 a k_j jsou skutečné napozorované hodnoty jejich počtů. Pro porovnání využijeme dvou kritérií. Označme $SS_e = \sum_{j=1}^{n^{pred}} (k_j - \lambda_j^{pred})^2$ součet čtverců odchylek predikce od skutečné hodnoty a $Abs_e = \sum_{j=1}^{n^{pred}} |k_j - \lambda_j^{pred}|$ součet absolutních odchylek. V tabulce 4.8 jsou porovnány predikční schopnosti jednotlivých modelů pro rok 2011 na základě těchto charakteristik (hodnoty jsou zde seřazeny od nejmenší po největší). Jak je vidět, hodnoty SS_e jsou ve všech případech přibližně stejné. Mírné rozdíly se projeví teprve porovnáním hodnot Abs_e. Jako nejpřesnější se z tohoto pohledu jeví Poissonův model, srovnatelné jsou pak hodnoty modelů pro panelová daty s tím, že jako nejvhodnější se jeví korelační struktura exchangeable (stejné korelace), a nakonec nejnižší kvalitu z tohoto pohledu vykazuje negativně binomický model.

riziková charakteristika	kategorie	odhad regresních koeficientů	odhad směrodatné odchylky	testová statistika	P-hodnota
intercept		0	0		
ts	1	-2,5308	0,0092	-275,04	<0,0001
ts	2	-2,3487	0,0081	-289,45	<0,0001
ts	3	-2,2388	0,0086	-259,66	<0,0001
ts	4	-2,1082	0,0226	-93,15	<0,0001
reg	1	0,2796	0,0091	30,78	<0,0001
reg	2	0,1610	0,0096	17,04	<0,0001
reg	3	0	0		
sv	0	-0,6505	0,0337	-19,32	<0,0001
sv	1	-0,0671	0,0081	-8,32	<0,0001
sv	2	0	0		
vek	1	0,5367	0,0197	27,23	<0,0001
vek	2	0,2128	0,0111	19,09	<0,0001
vek	3	0,0487	0,0089	5,47	<0,0001
vek	4	0	0		

Tabulka 4.7: Odhad na základě GEE metody.

	SS_e	Abs_e
Poissonův model	25858,84	13601,47
GEE- stejné korelace (exchangeable)	25860,25	13602,02
GEE- různé korelace (unstructured)	25864,73	13602,04
GEE- pásové korelace řádu 2 (2-dependance)	25865,51	13602,08
GEE- autoregrese řádu 1 (AR(1))	25869,06	13602,10
GEE- pásové korelace řádu 1 (1-dependance)	25869,16	13602,10
Negativně binomický model	25931,75	13602,38

Tabulka 4.8: Porovnání predikčních schopností.

Závěr

Zrekapitulujme si klíčové prvky této práce. V úvodu byly odvozeny základní vlastnosti hustot exponenciálního typu, načež bylo plynně navázáno formulací zobecněného lineární modelu. Dále jsme věnovali pozornost aplikaci teorie maximální věrohodnosti, speciálně byl odvozen explicitní tvar skórového vektoru a Fisherovy informační matice a navíc byla také představena numerická metoda iterativních nejmenších čtverců. V závěru první kapitoly byla formulována věta týkající se asymptotických vlastností.

Na počátku druhé kapitoly jsme představili poissonovskou regresi jako speciální případ zobecněného lineárního modelu v souvislosti s její možnou aplikací v pojišťovnictví. Ukázali jsme, jakým způsobem lze zohlednit expozice v riziku a interpretovat odhady parametrů a věrohodnostních rovnic. Ve zbylých sekcích byly uvedeny některé testy pro identifikaci významných regresorů a poté také možná využití reziduí při verifikaci modelu.

Třetí kapitola byla věnována alternativám poissonovské regrese. Na počátku jsme se zabývali diskuzí předpokladů nutných pro to, aby byly zachovány asymptotické vlastnosti odhadu. Bylo zde poznamenáno, že různá rizikovost pojistných smluv může způsobit zvýšení rozptylu v tarifních třídách a následně byl představen model negativně binomický, který je schopen tuto skutečnost reflektovat. Na závěr byla popsána GEE metoda, kterou lze aplikovat na panelová data, neboť na rozdíl od výše zmíněných přístupů nevyžaduje nezávislost odezev.

Nakonec byla provedena numerická studie na základě portfolia smluv povinného ručení osobních automobilů. Prezentovali jsme zde výsledky získané pomocí statistického softwaru SAS. Na základě skupiny regresorů, které byly identifikovány jako významné, bylo sestaveno několik modelů. Ty jsme pak porovnali z hlediska predikční schopnosti a jako nejkvalitnější se ukázal být model poissonovské regrese. Z odhadů pomocí GEE metody bylo nejlepšího výsledku dosaženo při volbě struktury stejných korelací mezi závisle proměnnými.

Příloha A

Metoda maximální věrohodnosti

Formulace níže uvedených definic a vět byla převzata z [1], přičemž některé doplňující poznatky byly čerpány z [12] a [8].

A.1 Základní definice a věty

Definice A.1. *Nechť náhodný vektor* $\mathbf{Y} = (Y_1, ..., Y_n)^T$ *má hustotou* $f(\mathbf{y}, \boldsymbol{\theta})$*, kde* $\boldsymbol{\theta}$ *je prvkem parametrického prostoru* Θ*. Při fixované hodnotě* $\mathbf{y}$ *se funkce* $f(\mathbf{y}, \boldsymbol{\theta}) = \ell(\boldsymbol{\theta})$ *nazývá věrohodnostní funkce a funkce* $\ln f(\mathbf{y}, \boldsymbol{\theta}) = L(\boldsymbol{\theta})$ *logaritmická věrohodnostní funkce. Hodnota* $\hat{\boldsymbol{\theta}}$ *parametru* $\boldsymbol{\theta}$*, která maximalizuje věrohodnostní funkci* $f(\mathbf{y}, \boldsymbol{\theta})$ *pro realizaci náhodného vektoru* $\mathbf{Y} = \mathbf{y}$*, se nazývá maximálně věrohodný odhad parametru* $\boldsymbol{\theta}$*.*

Poznámka. Dále se budeme zabývat zejména případy, kdy jsou náhodné veličiny $Y_1, ..., Y_n$ nezávislé, a tedy hustota náhodného vektoru $\mathbf{Y}$ je součinem marginálních hustot. V tomto případě je výhodné hledat hodnotu $\hat{\boldsymbol{\theta}}$ pomocí funkce $L(\boldsymbol{\theta})$. Je-li $\Theta \subset \mathbb{R}^m$, pak soustavu rovnic

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} = 0, \qquad i = 1, \ldots, m,$$

pomocí níž hledáme $\hat{\boldsymbol{\theta}}$, budeme nazývat soustavou věrohodnostních rovnic.

Definice A.2. *Nechť náhodný vektor* $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ *má hustotu* $f(\mathbf{y}, \boldsymbol{\theta})$ *vzhledem k nějaké* σ*-konečné míře* μ*. Řekneme, že systém takovýchto hustot* $\mathcal{F} = \{f(\mathbf{y}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ *je regulární, jsou-li splněny následující předpoklady:*

1. $\boldsymbol{\theta} \in \Theta$*, kde* Θ *je neprázdná otevřená množina v* $\mathbb{R}^m$*.*
2. *Množina* $M = \{\mathbf{y} : f(\mathbf{y}, \boldsymbol{\theta}) > 0\}$ *nezávisí na* $\boldsymbol{\theta}$ *.*
3. *Pro skoro všechna* $\mathbf{y} \in M$ *vzhledem k* μ *a pro všechna* $i = 1, \ldots, m$ *existují parciální derivace* $f_i'(\mathbf{y}, \boldsymbol{\theta}) = \frac{\partial f(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_i}$*.*
4. *Pro každé* i *a pro všechna* $\boldsymbol{\theta} \in \Theta$ *platí* $\int_M f_i'(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\mu(\mathbf{y}) = 0$*.*

5. Pro každou dvojici (i, j) existuje konečný integrál

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \int_M \frac{f'_i(\mathbf{y}, \boldsymbol{\theta}) f'_j(\mathbf{y}, \boldsymbol{\theta})}{f^2(\mathbf{y}, \boldsymbol{\theta})} f(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\mu(\mathbf{y}).$$

6. Matice $\mathcal{I}_n(\boldsymbol{\theta}) = (\mathcal{I}_{ij}(\boldsymbol{\theta}))_{i,j=1}^m$ je pozitivně definitní pro každé $\boldsymbol{\theta} \in \Theta$.

Matice $\mathcal{I}_n(\theta)$ se nazývá Fisherova informační matice. Hustota $f(\mathbf{y}, \boldsymbol{\theta})$ pocházející z regulárního systému hustot se pro stručnost označuje jako regulární.

Předpoklad 4 lze interpretovat jako požadavek na záměnnost integrálu a derivace, jelikož $\int_M f(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\mu(\mathbf{y}) = 1$. Vzorec pro $\mathcal{I}_{ij}(\boldsymbol{\theta})$ je také možné chápat jako integrál funkce

$$\frac{f'_i f'_j}{f^2} \tag{A.1.1}$$

dle míry s hustotou f vzhledem k μ. Aniž by došlo ke změně hodnoty $\mathcal{I}_{ij}(\boldsymbol{\theta})$, můžeme funkci (A.1.1) na doplňku množiny M libovolně dodefinovat a jako integrační obor vzít celý prostor $\mathbb{R}^n$. Využitím věty o přenosu integrace lze $\mathcal{I}_{ij}(\boldsymbol{\theta})$ vyjádřit ve tvaru střední hodnoty, tj.

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{f'_i(\mathbf{Y}, \boldsymbol{\theta}) f'_j(\mathbf{Y}, \boldsymbol{\theta})}{f^2(\mathbf{Y}, \boldsymbol{\theta})}\right] = \mathrm{E}\left[\frac{\partial \ln f(\mathbf{Y}, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(\mathbf{Y}, \boldsymbol{\theta})}{\partial \theta_j}\right]. \tag{A.1.2}$$

Je-li $\mathbf{Y} = (Y_1, ..., Y_n)^T$ vektor nezávislých stejně rozdělených veličin, pak platí $\mathcal{I}_n(\boldsymbol{\theta}) = n\mathcal{I}_1(\boldsymbol{\theta})$ (viz [1] věta 7.20). Dále v textu budeme pro zjednodušení užívat symbol $\mathcal{I}(\boldsymbol{\theta})$ namísto $\mathcal{I}_1(\boldsymbol{\theta})$.

Definice A.3. *Mějme náhodný vektor $\mathbf{Y} = (Y_1, ..., Y_n)^T$ s regulární hustotou $f(\mathbf{y}, \boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T \in \Theta \subset \mathbb{R}^m$. Pak m-rozměrný náhodný vektor*

$$U = U(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), \ldots, U_m(\boldsymbol{\theta}))^T = \frac{\partial \ln f(\mathbf{Y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \ln f(\mathbf{Y}, \boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ln f(\mathbf{Y}, \boldsymbol{\theta})}{\partial \theta_m}\right)^T$$

se nazývá skórový vektor.

Věta A.4. *Nechť je systém hustot $\{f(\mathbf{y}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ regulární. Předpokládejme navíc, že pro skoro všechna $\mathbf{y} \in M = \{\mathbf{y} : f(\mathbf{y}, \boldsymbol{\theta}) > 0\}$ (vzhledem k μ) existují derivace*

$$f''_{ij}(\mathbf{y}, \boldsymbol{\theta}) = \frac{\partial^2 f(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}, \qquad i, j = 1, \ldots, m,$$

a že pro všechna $\boldsymbol{\theta} \in \Theta$ platí

$$\int_M f''_{ij}(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\mu(\mathbf{y}) = 0, \qquad i, j = 1, \ldots, m. \tag{A.1.3}$$

Pak platí

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = -\int_M \frac{\partial^2 \ln f(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} f(\mathbf{y}, \boldsymbol{\theta}) \mathrm{d}\mu(\mathbf{y}), \qquad i, j = 1, \ldots, m. \tag{A.1.4}$$

Důkaz. Derivováním dostaneme vztah

$$\frac{\partial^2 \ln f}{\partial\theta_i \partial\theta_j} = \frac{f''_{ij} f - f'_i f'_j}{f^2} = \frac{f''_{ij}}{f} - \frac{f'_i f'_j}{f^2}, \tag{A.1.5}$$

který platí pro skoro všechna $x \in M$. Využitím (A.1.3) a (A.1.5) dostáváme

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \int_M \left(\frac{f'_i f'_j}{f^2}\right) f \mathrm{d}\mu$$
$$= \int_M \left(\frac{f''_{ij}}{f}\right) f \mathrm{d}\mu - \int_M \left(\frac{\partial^2 \ln f}{\partial\theta_i \partial\theta_j}\right) f \mathrm{d}\mu = -\int_M \left(\frac{\partial^2 \ln f}{\partial\theta_i \partial\theta_j}\right) f \mathrm{d}\mu.$$

□

Shrnutím dosavadních poznatků můžeme získat hned několik možných vyjádření Fisherovy informační matice. Při splnění předpokladů věty A.4 totiž pro složky $\mathcal{I}_n(\theta)$ zřejmě platí:

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \mathrm{E}\,\frac{f'_i(\mathbf{y},\boldsymbol{\theta}) f'_j(\mathbf{y},\boldsymbol{\theta})}{f^2(\mathbf{y},\boldsymbol{\theta})} = -\mathrm{E}\,\frac{\partial^2 \ln f(\mathbf{y},\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} = -\mathrm{E}\,\frac{\partial U_i(\boldsymbol{\theta})}{\partial\theta_j}.$$

Někdy se též používá maticové značení:

$$\mathcal{I}_n(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right] = -\mathrm{E}\left[\frac{\partial \ln f(\mathbf{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right].$$

Další možné vyjádření plyne z (A.1.2) a z podmínky 4 v definici (A.2):

$$\mathcal{I}_n(\boldsymbol{\theta}) = \mathrm{E}[U(\boldsymbol{\theta})U(\boldsymbol{\theta})^T] = \mathrm{var}\, U(\boldsymbol{\theta}). \tag{A.1.6}$$

Poznamenejme, že zároveň z podmínky 4 v definici systému regulárních hustot (viz (A.2)) díky vyjádření

$$\int_M f'_i(\mathbf{y},\boldsymbol{\theta}) \mathrm{d}\mu(\mathbf{y}) = \mathrm{E}\,\frac{f'_i(\mathbf{y},\boldsymbol{\theta})}{f(\mathbf{y},\boldsymbol{\theta})} = \mathrm{E}\,\frac{\partial \ln f(\mathbf{Y},\boldsymbol{\theta})}{\partial\theta_i}$$

také plyne

$$\mathrm{E}\, U(\boldsymbol{\theta}) = 0.$$

A.2 Vlastnosti maximálně věrohodných odhadů

Důvodem hojného užívání maximálně věrohodných odhadů je především jejich chování v případě dostatečného velkého datového vzorku. Bohužel takovéto odhady nejsou obecně nestranné a zpravidla nelze říci, při jakém počtu pozorování je již možné se spolehnout na jejich asymptotické vlastnosti. V pojišťovnictví se ovšem pracuje téměř výhradně s rozsáhlými datovými soubory, proto se s takovýmto problémem setkáme pouze velmi zřídka. Zjednodušeně lze říci, že odhady metodou maximální věrohodnosti mají následující vlastnosti:

- Konzistence
- Asymptotická normalita
- Asymptotická eficience (nejnižší možný rozptyl)

V této je sekci blíže vysvětlen význam těchto vlastností a jsou zde uvedeny postačující předpoklady pro jejich platnost.

A.2.1 Konzistence a asymptotická normalita

Věta A.5. *Mějme dán regulární systém hustot* $\{f(\mathbf{y}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ *s Fisherovou informační maticí* $\mathcal{I}(\boldsymbol{\theta})$*. Nechť jsou navíc splněny následující předpoklady.*

1. *Parametrický prostor* $\Theta \subset \mathbb{R}^m$ *obsahuje otevřený interval* ω*, v němž se nachází skutečná hodnota parametru* $\boldsymbol{\theta} = \boldsymbol{\theta_0}$*.*
2. *Náhodný vektor* $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$*, tvoří nezávislé a stejně rozdělené veličiny s hustotou* $f(y, \boldsymbol{\theta})$ *vzhledem k nějaké* σ*-konečné míře* μ*.*
3. *Množina* $M = \{y : f(y, \boldsymbol{\theta}) > 0\}$ *nezávisí na* $\boldsymbol{\theta}$*.*
4. *Nechť* $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \Theta$*. Pak* $f(y, \boldsymbol{\theta_1}) = f(y, \boldsymbol{\theta_2})$ *skoro jistě vzhledem k* μ *právě tehdy, když* $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$*.*
5. *Derivace* $\frac{\partial^3 f(y,\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_k}$ *existuje pro skoro všechna* y*, pro všechna* $\boldsymbol{\theta} \in \omega$ *a pro všechna* $i, j, k = 1, \ldots, m$*.*
6. *Pro všechna* $\boldsymbol{\theta} \in \omega$ *platí*
$$\int_M f''_{ij}(y, \boldsymbol{\theta}) \mathrm{d}\mu(y) = 0, \qquad i, j = 1, \ldots, m.$$
7. *Pro všechna* $i, j, k = 1, \ldots, m$ *existují funkce* $M_{ijk}(y) \geq 0$ *takové, že* $E_{\theta_0} M_{ijk}(Y) < \infty$ *a* $\mid \frac{\partial^3 \ln f(y,\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_k} \mid \leq M_{ijk}(y)$ *pro všechna* $\boldsymbol{\theta} \in \omega$ *a skoro všechna* $y \in M$*.*

Pak má maximálně věrohodný odhad následující vlastnosti:

(i) Pokud $n \to \infty$*, pak pro každé* $\varepsilon > 0$ *existuje s pravděpodobností blížící se jedné takové řešení* $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ *soustavy věrohodnostních rovnic, že* $\parallel \hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0} \parallel < \varepsilon$*.*

(ii) Pro $n \to \infty$ *platí*
$$\frac{1}{\sqrt{n}} U(\boldsymbol{\theta_0}) \xrightarrow{d} N(0, [\mathcal{I}(\boldsymbol{\theta_0})]).$$

(iii) Existuje-li pro každé dostatečně velké n *a pro každou hodnotu* $\mathbf{Y}$ *takový kořen* $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ *systému věrohodnostních rovnic, že* $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ *je konzistentním odhadem parametru* $\boldsymbol{\theta_0}$*, pak*
$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}) \xrightarrow{d} N(0, [\mathcal{I}^{-1}(\boldsymbol{\theta_0})]). \tag{A.2.1}$$

Důkaz. Podrobně je důkaz věty proveden v knize [12] v kapitole 6.5 (Theorem 5.1). □

Poznámka. Větu lze formulovat i pro případ nezávislých stejně rozdělených vektorů, aniž by bylo nutné zásadně měnit její předpoklady (viz [9]). Existují ovšem i silnější zobecnění věty, které navíc nevyžadují, aby skutečná hustota Y_i měla tvar $f(y, \boldsymbol{\theta})$ pro nějaké $\boldsymbol{\theta} \in \Theta$ (viz [17]). Jiné formulace zase nepředpokládají nezávislé a stejně rozdělené náhodné veličiny. Tato zobecnění byla vyvinuta pro účely zobecněných lineárních modelů (viz věta (1.6)) a jejich rozšíření, ale také například časových řad, jejich přesné formulace lze vyhledat v odborných článcích (např. [2, 6, 7]).

A.2.2 Eficience

Asymptotická eficience

Věta A.6. *Mějme regulární systém hustot* $\{f(\mathbf{y}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\}$. *Nechť* $\mathbf{T}(\mathbf{Y}) = \mathbf{T} = (T_1, \ldots, T_m)^T$ *je nestranný odhad* $\boldsymbol{\theta}$ *takový, že* $\mathrm{E}T_i^2 < \infty$ *pro každé* $i = 1, \ldots, m$ *a pro každé* $\boldsymbol{\theta} \subset \Theta$. *Dále nechť platí*

$$\int_M T_i(y) \frac{\partial f(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_j} \mathrm{d}\mu(y) = \delta_{ij}, \qquad i, j = 1, \ldots, m,$$

pak dostáváme

$$\operatorname{var} \mathbf{T} - \mathcal{I}_n^{-1}(\boldsymbol{\theta}) \geq 0.$$

Důkaz. Důkaz je uveden v [1] (věta 7.31). □

Věta pouze ukazuje, jakým způsobem je možné zdola omezit velikost rozptylu. Nestranné odhady, které mají rozptyl roven inverzní hodnotě Fisherovy informační matice $\mathcal{I}_n^{-1}(\boldsymbol{\theta})$, nazýváme eficientní a říkáme, že dosahují tzv. dolní Raovy-Cramerovy meze.

Navíc lze předchozí větu dále zobecnit pro třídu asymptoticky nestranných odhadů. Rozlišit kvalitu konzistentních odhadů lze na základě asymptotické hodnoty $\sqrt{n}\operatorname{var} \mathbf{T}$. Za předpokladu asymptotické normality $\sqrt{n}\mathbf{T}$ existuje také pro matici $lim_{n\to\infty}\sqrt{n}\operatorname{var} \mathbf{T}$ jistá dolní mez, která je rovna hodnotě $\mathcal{I}^{-1}(\boldsymbol{\theta})$. Z vlastnosti (iii) uvedené ve větě (A.5) plyne, že maximálně věrohodné odhady jsou asymptoticky eficientní, tj. nejlepší ve smyslu neexistence asymptoticky normálního konzistentního odhadu $\mathbf{T}$, pro který by platilo $lim_{n\to\infty}\sqrt{n}\operatorname{var} \mathbf{T} \leq \mathcal{I}^{-1}(\boldsymbol{\theta})$. Korektně je problematika asymptoticky eficientních odhadu popsána v kapitole 6 knihy [12].

A.3 Testování hypotéz

Jedním z aspektů kvalitního modelu je kromě schopnosti vysvětlit závislost sledované veličiny na regresorech také jeho jednoduchost a snadná interpretovatelnost. Mezi dvěma modely s přibližně stejnými vlastnostmi obvykle preferujeme ten s menším počtem nezávisle proměnných. Pro identifikaci toho, zda daná skupinka proměnných

významně napomáhá vysvětlení závislosti slouží různé statistické testy vycházející z vlastností maximálně věrohodných odhadů, případně se přihlíží k hodnotám informačních či jiných kritérií. V této sekci jsou blíže popsány příklady takovýchto statistických testů.

Věta A.7. *Nechť je matice $\mathcal{I}(\boldsymbol{\theta})$ spojitá v bodě $\boldsymbol{\theta_0}$ a $\hat{\mathcal{I}}_n$ konzistentní odhad $\mathcal{I}(\boldsymbol{\theta_0})$, při $n \to \infty$. Pak za předpokladů uvedených ve větě A.5 mají statistiky:*

$$\begin{aligned} LR(\boldsymbol{\theta_0}) &= 2[L(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) - L(\boldsymbol{\theta_0})] \\ W(\boldsymbol{\theta_0}) &= n(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0})^T \hat{\mathcal{I}}_n (\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}) \\ LM(\boldsymbol{\theta_0}) &= \frac{1}{n}[U(\boldsymbol{\theta_0})]^T \hat{\mathcal{I}}_n{}^{-1} [U(\boldsymbol{\theta_0})] \end{aligned}$$

asymptoticky χ^2_m rozdělení.

Důkaz. Viz [1] poznámka 8.17. □

Na základě věty A.7 můžeme testovat hypotézu $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$ proti alternativě $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta_0}$. Testy založené na $LR(\boldsymbol{\theta_0})$, $W(\boldsymbol{\theta_0})$, $LM(\boldsymbol{\theta_0})$ se po řadě nazývají: test poměrem věrohodností (likelihood ratio), Waldův test a skórový test (Lagrange multiplier). Skórovému testu se také někdy říká Raův nebo méně často test založený na Lagrangerových multiplikátorech. Nulovou hypotézu všech tří výše zmíněných testů zamítáme, překročí-li testová statistika kritickou hodnotu $\chi^2_m(1-\alpha)$, tj., $[(1-\alpha)100]\%$ kvantil.

Označíme

$$F_i(\boldsymbol{\theta}) = \left(-\frac{\partial^2 \ln f(Y_i, \boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} \right)_{r,s=1}^{m}, \qquad F(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} F_i(\boldsymbol{\theta}) \tag{A.3.1}$$

Z (A.1.4) plyne, že $\mathrm{E}F_i(\boldsymbol{\theta_0}) = \mathcal{I}(\boldsymbol{\theta_0})$, a proto je také zřejmě $\mathrm{E}F(\boldsymbol{\theta_0}) = \mathcal{I}(\boldsymbol{\theta_0})$ a $F(\boldsymbol{\theta_0}) \xrightarrow{P} \mathcal{I}(\boldsymbol{\theta_0})$, tj., $F(\boldsymbol{\theta_0})$ je nestranným a konzistentním odhadem $\mathcal{I}(\boldsymbol{\theta_0})$. Z tohoto důvodu se $F(\boldsymbol{\theta})$ nazývá *výběrovou Fisherovou informační maticí.*

Lemma A.8. Mějme regulární matici

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}$$

se čtvercovými bloky $\mathcal{I}_{11}$ a $\mathcal{I}_{22}$. Dále označme

$$\begin{aligned} \mathcal{I}_{11.2} &= \mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21} \\ \mathcal{I}_{22.1} &= \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}. \end{aligned}$$

Pak

$$\mathcal{I}^{-1} = \begin{pmatrix} \mathcal{I}_{11.2}^{-1} & -\mathcal{I}_{11.2}^{-1}\mathcal{I}_{12}\mathcal{I}_{22}^{-1} \\ -\mathcal{I}_{22.1}^{-1}\mathcal{I}_{21}I_{11}^{-1} & \mathcal{I}_{22.1}^{-1} \end{pmatrix}.$$

Pro další potřeby této sekce předpokládejme, že m-rozměrný vektor parametrů má tvar

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta_{(1)}} \\ \boldsymbol{\theta_{(2)}} \end{pmatrix} \begin{matrix} \}k \\ \}m-k \end{matrix} \qquad , m \geq 2, 1 \leq k < m.$$

Naším cílem je testovat hypotézu $H_0 : \boldsymbol{\theta_{(1)}} = \boldsymbol{\theta_{(1)0}}$ proti alternativě $H_1 : \boldsymbol{\theta_{(1)}} \neq \boldsymbol{\theta_{(1)0}}$. Nejčastěji je $\boldsymbol{\theta_{(1)0}} = \mathbf{0}$, neboť pak nulová hypotéza odpovídá platnosti podmodelu s $m-k$ parametry a alternativa platnosti plného modelu obsahujícího m parametrů. Dále položme

$$U(\boldsymbol{\theta}) = \begin{pmatrix} U_1(\boldsymbol{\theta}) \\ U_2(\boldsymbol{\theta}) \end{pmatrix} \begin{matrix} \}k \\ \}m-k, \end{matrix}$$

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{I}_{11}(\boldsymbol{\theta}) & \mathcal{I}_{12}(\boldsymbol{\theta}) \\ \mathcal{I}_{21}(\boldsymbol{\theta}) & \mathcal{I}_{22}(\boldsymbol{\theta}) \end{pmatrix} \begin{matrix} \}k \\ \}m-k, \end{matrix}$$

kde $\mathcal{I}_{11}(\boldsymbol{\theta})$ je čtvercová matice typu $k \times k$. Maximálně věrohodný odhad za platnosti hypotézy $H_0 : \boldsymbol{\theta_{(1)}} = \boldsymbol{\theta_{(1)0}}$ označme

$$\tilde{\boldsymbol{\theta}}_n = \begin{pmatrix} \boldsymbol{\theta_{(1)0}} \\ \tilde{\boldsymbol{\theta}}_{\mathbf{(2)0}} \end{pmatrix},$$

a standardní odhad odpovídající alternativě

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} = \begin{pmatrix} \hat{\boldsymbol{\theta}}_{\mathbf{(1)}\boldsymbol{n}} \\ \hat{\boldsymbol{\theta}}_{\mathbf{(2)}\boldsymbol{n}} \end{pmatrix}.$$

Věta A.9. *Nechť jsou splněny předpoklady věty (A.5) a matice $\mathcal{I}(\boldsymbol{\theta})$ je spojitá v bodě $\boldsymbol{\theta_0}$ odpovídajícímu skutečné hodnotě parametru. Nechť navíc $\hat{\mathcal{I}}_{11.2n}$ je konzistentní odhad $\mathcal{I}_{11.2}(\boldsymbol{\theta_0})$, pak za platnosti hypotézy $H_0 : \boldsymbol{\theta_{(1)}} = \boldsymbol{\theta_{(1)0}}$ mají statistiky*

$$\begin{aligned} LM^* &= \frac{1}{n}[U_1(\tilde{\boldsymbol{\theta}}_n)]^T[\hat{\mathcal{I}}_{11.2n}]^{-1}[U(\tilde{\boldsymbol{\theta}}_n)] \\ W^* &= n(\hat{\boldsymbol{\theta}}_{\mathbf{(1)}\boldsymbol{n}} - \boldsymbol{\theta_{(1)0}})^T\hat{\mathcal{I}}_{11.2n}(\hat{\boldsymbol{\theta}}_{\mathbf{(1)}\boldsymbol{n}} - \boldsymbol{\theta_{(1)0}}) \qquad \text{(A.3.2)} \\ LR^* &= 2[L(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) - L(\tilde{\boldsymbol{\theta}}_n)] \qquad \text{(A.3.3)} \end{aligned}$$

asymptoticky χ^2_k rozdělení.

Důkaz. Viz [1] věta 8.25. □

Poznámka. Jako $\hat{\mathcal{I}}_{11.2n}$ se nejčastěji volí výběrová Fisherova informační matice (viz A.3.1). Tvar $F(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}})$ se užívá v případě Waldova testu a $F(\tilde{\boldsymbol{\theta}}_n)$ pokud jde o skórový test. Asymptoticky mají všechny uvedené testy stejnou sílu, avšak především pro menší počty pozorování preferujeme z dle [9] obvykle test poměrem věrohodností.

Příloha B

Vybrané části zdrojového kódu

V nadcházejících zdrojových kódech bylo užito následujícího značení:

`selekce_2010_vyb`	název datového souboru A (2010),
`selekce_2008_2010`	název datového soubor B (2008-2010),
`ts`	tarifní skupina (dle objemu motoru),
`reg`	počet obyvatel v místě bydliště pojistníka,
`sv`	stáří vozidla,
`vek`	věk pojištěného,
`ppuM`	počet pojistných událostí nepřesahujících 500 000 Kč.,
`logd`	offset (logaritmus expozice v riziku),
`smlo_pk`	klíč jednoznačně určující smlouvu.

Poissonův model

```
ods output ParameterEstimates=work.estimates;
proc genmod data=selekce_2010_vyb;
   class ppuM ts reg sv/*(ref="0")*/ vek;
   model ppuM = ts reg sv vek /
   noint
   dist=poisson
   offset= logd
   link=log
   type1 type3;
   output out = Residuals_pois
          pred = Pred
          STDRESDEV = stdresdev;
quit;
```

ODP model

```
ods output ParameterEstimates=work.estimates_ODP;
proc genmod data=selekce_2010_vyb;
   class ppuM ts reg sv vek;
   model ppuM = ts reg sv vek /
   noint
   dist=poisson
   offset= logd
   SCALE=PEARSON
   link=log
   type1 type3;
quit;
```

Robustní odhad rozptylu

```
proc genmod data=selekce_2010_vyb;
   class ppuM ts reg sv vek smlo_pk;
   model ppuM = ts reg sv vek /
   noint
   dist=poisson
   offset= logd
   link=log
   type1 type3;
   repeated subject=smlo_pk / type=ind;
quit;
```

Negativně binomický model

```
ods output ParameterEstimates=work.estimates_negbin;
proc genmod data=selekce_2010_vyb;
   class ppuM ts reg sv vek;
   model ppuM = ts reg sv vek /
   noint
   dist=negbin
   offset= logd
   link=log
   type1 type3;
quit;
```

GEE metoda

```
ods output GEEEmpPEst =work.estimates_GEE;
proc genmod data=selekce_2008_2010;
   class ppuM ts reg sv vek smlo_pk;
   model ppuM = ts reg sv vek /
   noint
   dist=poisson
   offset= logd
   link=log
   type1 type3;
   repeated subject=smlo_pk / type=UNSTR covb corrw;
quit;
```

Literatura

[1] Anděl, J. *Základy matematické statistiky*. Matfyzpress, Praha 2007. ISBN 80-7378-001-1

[2] Bradley, R. A. a Gart J. J. The Asymptotic Properties of ML Estimators when Sampling from Associated Populations. *Biometrika*, Vol. 49, No. 1/2 (1962), pp. 205–214

[3] Branda, M. *Tvorba optimálních sazeb v neživotním pojištění* [online]. Prezentace k semináři z aktuárských věd, 2013. <http://artax.karlin.mff.cuni.cz/~branm1am/mbvyuka.html>

[4] Denuit, M., Maréchal, X., Pitrebois, S., Walhin J. *Actuarial Modelling of Claim Count*. John Wiley & Sons, Ltd, England 2007. ISBN 978-0-470-02677-9

[5] Ohlsson, E a Johansson, B. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, Berlin Heidelberg 2010. ISBN 978-3-642-10790-0

[6] Fahrmeir, L a Kaufmann, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models *The Annals of Mathematical Statistics*, Vol. 13, No. 1 (1985), pp. 342–368.

[7] Hoadley, B. Asymptotic Properties of Maximum Likelihood Estimators for the Independent Not Identically Distributed Case. *The Annals of Mathematical Statistics*, Vol. 42, No. 6 (1971), pp. 1977–1991

[8] Hrdličková, Z. *Mnohorozměrné zobecněné lineární modely*. Dizertační práce. Brno,2006. KAM PřF MU. vedoucí doc. RNDr. Jaroslav Michálek, CSc.

[9] Kulich, M. Zápisky z přednášky: Zobecněné lineární modely (NSTP196), 2012

[10] Liang, K. a Scott, L. Z. Longitudinal data analysis using generalized linear models. *Biometrika*, Vol. 73, No. 1 (1986), pp. 13–22

[11] Lachout, P. *Matematické programování*. pracovní text k přednášce optimalizace I, 2008

[12] Lehmann, E a George, C. *Theory of point estimation*. Springer, New York 1998. 2nd ed.

[13] McCullagh, P a Nelder, J.A. *Generalized Linear Models*. Chapman and Hall/CRC, 1999. 2nd. ed.

[14] Pierce, D. A. a Schafer D. W. Residuals in Generalized Linear Models. *Journal of the American Statistical Association*, Vol. 81, No.396 (1986), pp. 977–986

[15] Piet, D. J. a Heller, G. Z. *Generalized Linear Models for Insurance Data*. Cambridge University Press, 2008. ISBN-13 978-0-511-38677-0

[16] SAS Institute Inc., 2008. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc., chapter 37 (The GENMOD Procedure)

[17] White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*. Vol. 50, No. 1 (Jan., 1982), pp. 1–26

Seznam použitých zkratek

$\mathbb{N}_0$ ..... množina nezáporných celých čísel
$\mathbb{R}$ ..... množina reálných čísel
$\mathbb{R}^n$ ..... reálný n-rozměrný euklidovský prostor
$\mathbb{R}^+$ ..... $\{x \in \mathbb{R} : x > 0\}$
p-rozměrný vektor $\mathbf{x}$ ..... p-rozměrný sloupcový vektor $\mathbf{x}$
matice $\mathbf{C}^T$ ..... transponovaná matice $\mathbf{C}$
$\mathbf{C} \geq 0$ ..... $\mathbf{C}$ je pozitivně definitní.
$diag(w_1, \ldots, w_n)$ ..... diagonální matice $\begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix}$
$\| \mathbf{x} \|$ ..... euklidovská norma vektoru $\mathbf{x}$
$\| \mathbf{C} \|$ ..... norma čtvercové matice souhlasná (compatible) s euklidovskou normou
Tj., je-li $\mathbf{C}$ typu $p \times p$ a $\mathbf{x}$ vektor délky p, pak platí $\| \mathbf{Cx} \| \leq \| \mathbf{C} \| \cdot \| \mathbf{x} \|$.
$b \in C^n(\Theta)$ ..... Funkce b je n-krát spojitě diferencovatelná na Θ.
$\frac{\partial f}{\partial \boldsymbol{\theta}}$ ..... derivace skalární funkce f podle vektorového parametru $\boldsymbol{\theta}$.
Tj. je-li $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$, pak $\frac{\partial f}{\partial \boldsymbol{\theta}} = \left(\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_m} \right)^T = (f'_1, \ldots, f'_m)^T$.
$\frac{\partial f}{\partial \boldsymbol{\theta}^T}$ ..... derivace skalární funkce f podle vektorového parametru $\boldsymbol{\theta}^T$.
Tj. je-li $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_m)$, pak $\frac{\partial f}{\partial \boldsymbol{\theta}^T} = \left(\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_m} \right) = (f'_1, \ldots, f'_m)$.
$\frac{\partial U}{\partial \boldsymbol{\theta}^T}$ ..... derivace vektorové funkce U podle vektorového parametru $\boldsymbol{\theta}^T$.
Tj. je-li $U = (U_1, \ldots, U_n)^T$ a $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_m)$, pak

$$\frac{\partial U}{\partial \boldsymbol{\theta}^T} = \begin{pmatrix} \frac{\partial U_1}{\partial \boldsymbol{\theta}^T} \\ \vdots \\ \frac{\partial U_n}{\partial \boldsymbol{\theta}^T} \end{pmatrix} = \begin{pmatrix} \frac{\partial U_1}{\partial \theta_1} & \cdots & \frac{\partial U_1}{\partial \theta_m} \\ \vdots & & \vdots \\ \frac{\partial U_n}{\partial \theta_1} & \cdots & \frac{\partial U_n}{\partial \theta_m} \end{pmatrix}.$$

$\ll$ ..... řádově menší než
$\sim$ ..... Náhodný vektor (resp. veličina) má rozdělení
$\approx$ ..... Náhodný vektor (resp. veličina) má přibližně rozdělení
$\xrightarrow{P}$ ..... konvergence v pravděpodobnosti
$\xrightarrow{d}$ ..... konvergence v distibuci
χ^2_n ..... χ^2 rozdělení o n stupních volnosti
$N_p(\boldsymbol{\mu}, \Sigma)$ ..... p-rozměrné normální rozdělní se střední hodnotou $\boldsymbol{\mu}$ a rozptylovou maticí Σ
$\dim(\mathbf{Z})$ ..... počet složek náhodného vektoru $\mathbf{Z}$
$\mathrm{E}\,\mathbf{Y}$ a $\mathrm{var}\,\mathbf{Y}$ ..... střední hodnota a rozptylová matice náhodného vektoru $\mathbf{Y}$
$\mathrm{corr}\,(N_i, N_j)$ ..... korelace mezi náhodnými veličinami N_i a N_j

Tvary hustot známých rozdělní užitých v této práci, jsou uvedeny v podsekci 1.1.1.