

GLM MODEL PRIRÁŽOK A ZLIAV PRI OCEŇOVANÍ V NEŽIVOTNM POISTENÍ

Ján Gogola, Silvia Zelinová

Abstrakt

Na určovanie poistného v neživotnom poistení neexistuje univerzálna metóda ani metodika. Spôsoby výpočtu jednotlivých poistovní sa v technike kalkulovania poistného viac či menej líšia. Pri kalkulácii poistného sa vychádza buď z prognózovania celkového poistného v uvažovanej tarifnej skupine za určité poistné obdobie (najčastejšie jeden rok), alebo z prognózovania ukazovateľov, pomocou ktorých možno poistné vyjadriť.

V tomto príspevku sa venujeme teórii a implementácii zovšeobecnených lineárnych modelov v oblasti oceňovania v neživotnom poistení. Predstavíme niektoré najčastejšie používané metódy odhadu. V praktickej časti budeme demonštrovať použitie programu R na reálnych poistných dátach.

Kľúčové slová

GLM – zovšeobecnený lineárny model, oceňovanie v neživotnom poistení, multiplikatívny model

1. ÚVOD

V homogénnom portfóliu, kde majú všetci poistenci rovnakú mieru rizika, nie je dôvod mať variabilnú výšku poistného. V praxi je väčšina portfólií heterogénna. Miešajú jednotlivcov s rôznymi úrovňami rizika. Je nevyhnutná potreba klasifikácie rizík.

V súčasnosti je pre poisťovne mimoriadne ťažké udržať krížové dotácie medzi rôznymi kategóriami rizika v konkurenčnom prostredí. Poistní matematici preto musia navrhnúť tarifnú štruktúru, ktorá spravodlivo rozdelí bremeno nárokov medzi poistencov. Na dosiahnutie klasifikácie rizika použijeme zovšeobecnené lineárne modely (GLM).

Tvorba sadziieb (alebo klasifikácia rizika) je v podstate o klasifikácii poistných zmlúv podľa ich rizikových charakteristík.

Klasifikačné premenné alebo ratingové premenné sú tzv. apriórne premenné, pretože ich hodnoty je možné určiť skôr, ako poistník začne byť krytý poisťovňou.

V poistení zodpovednosti za škodu spôsobenú prevádzkou motorového vozidla (PZP MV) zahŕňajú vek, pohlavie a povolanie poistníka, typ a použitie jeho auta, miesto, kde býva, atď.

Tieto pozorovateľné charakteristiky sa zvyčajne považujú za nenáhodné ratingové premenné. Aj keď sú v cenníkoch zahrnuté všetky ratingové premenné, medzi jednotlivými vodičmi zostávajú značné rizikové rozdiely (kvôli skrytým vlastnostiam, ako je temperament a zručnosť, agresivita za volantom, znalosť diaľnice atď.)

2. LINEÁRNY REGRESNÝ MODEL

Lineárny regresný model tvorí základ takmer každej štatistickej analýzy a modelovania. V poisťovníctve sa najčastejšie modeluje očakávaný počet poistných udalostí, výška škody, ktorá nastane realizáciou poistného rizika a pod. Klasický lineárny regresný model v tomto prípade nestačí. Využívane sú miesto toho zložitejšie modely, známe pod názvom Zovšeobecnené lineárne modely, skrátene GLM (Generalized Linear Models). Klasický lineárny reg-

resný model tvorí základ pre pochopenie GLM, a preto si pripomeňme v nasledujúcej časti základnú metodológiu tohto modelu.

Klasický lineárny regresný model môžeme zapísať v tvare:

$$y_i = \mu_i + \varepsilon_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \dots + \beta_p \cdot x_{ip} + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

kde:

y_i je i -té pozorovanie náhodnej veličiny Y , pre $i = 1, \dots, n$,

x_{ij} je i -tá hodnota vysvetľujúcich premenných, pre $i = 1, \dots, n$ a $j = 1, \dots, p$,

β_j sú neznáme parametre modelu pre $j = 1, \dots, p$,

ε_i označuje i -tu náhodnú zložku modelu, pre $i = 1, \dots, n$.

Pre náhodnú zložku ε_i platia nasledujúce predpoklady:

- $E(\varepsilon_i) = 0$,
- homoskedasticita, tj. rozptyl náhodnej zložky je rovnaký pre všetky pozorovania $Var(\varepsilon_i) = \sigma^2 > 0$,
- $cov(\varepsilon_i, \varepsilon_j) = 0$, pre $i \neq j$ (tj. nekorelovanosť rôznych dvojíc pozorovaní vysvetľovanej premennej Y^1),
- normálne rozdelenie $\varepsilon_i \sim N(0; \sigma^2)$.

Z $\varepsilon_i \sim N(0; \sigma^2)$ vyplýva $y_i \sim N(\mu_i; \sigma^2)$.

V maticovom tvare je možné klasický lineárny regresný model zapísať ako:

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0; \sigma^2 \mathbf{I}), \quad (2)$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$, $\mathbf{y} = (y_1, y_2, \dots, y_n)'$, $\mathbf{I}$ je jednotková matica typu $n \times n$ a $\mathbf{X}^2$ je matica typu $n \times (p + 1)$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ & \vdots & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Odhad neznámych parametrov $\boldsymbol{\beta}$ dostávame metódou najmenších štvorcov. Model je lineárny v zmysle parametrov $\boldsymbol{\beta}$.

3. GLM – ZOVŠEOBECNENÝ LINEÁRNY MODEL

Zovšeobecnený lineárny model (GLM) sa využíva pre modelovanie a ohodnotenie vzťahu medzi cieľovou a vysvetľujúcimi premennými. Zovšeobecnený lineárny model je rozšírením klasického lineárneho regresného modelu, keď predpoklady klasického modelu nie sú splnené, čo v poisťovacej praxi je častý jav. Podstatným rozdielom oproti klasickému regresnému

¹ Náhodná veličina Y sa nazýva aj cieľová premenná alebo response

² $\mathbf{X}$ je rozšírená matica realizácií nezávislých premenných x_{ij}

modelovaniu spočíva v tom, že rozdelenie závislej premennej už nemusí byť normálne ale je z tzv. *triedy exponenciálnych rozdelení* (Exponential Dispersion Family).

Všeobecný tvar hustoty rozdelenia z triedy exponenciálnych rozdelení môžeme zapísať v tvare:

$$f(y; \theta, \phi) = c(y; b(\phi)) \cdot \exp \left\{ \frac{y\theta - a(\theta)}{b(\phi)} \right\}, \quad y \in R, \quad (3)$$

kde θ a ϕ sú neznáme parametre rozdelenia.

Parameter θ sa nazýva *kanonický parameter* a ϕ sa nazýva *disperzný parameter*.

Známe funkcie $a(\theta)$, $b(\phi)$ a $c(y; \phi)$ určujú pravdepodobnostnú funkciu rozdelenia.

Pre každý člen náhodnej veličiny y , ktorý má rozdelenie z triedy exponenciálnych rozdelení platí:

$$E(y) = a'(\theta), \quad \text{Var}(y) = b(\phi) \cdot a''(\theta)^3 = b(\phi) \cdot V(\mu). \quad (4)$$

Rozdiely GLM modelu oproti klasickému lineárnemu modelu sú nasledujúce:

- závislá premenná y má rozdelenie z triedy exponenciálnych rozdelení,
- lineárny prediktor⁴ $\eta_i = \beta_0 + \sum_{j=1}^p \beta_j \cdot x_{ij}$, pre $i = 1, \dots, n$,
- linková funkcia $g(\cdot)$ je monotónna, diferencovateľná funkcia a určuje vzťah strednej hodnoty k nezávislým premenným (k lineárnemu prediktoru)

$$\eta_i = g(\mu_i) \quad \text{alebo} \quad \mu_i = g^{-1}(\eta_i). \quad (5)$$

V nasledujúcej tabuľke uvádzame niektoré známe rozdelenia z triedy exponenciálnych rozdelení s ich hlavnými charakteristikami.

Rozdelenie	Rozsah y	Disperzný parameter ϕ	Kanonická link funkcia	Stredná hodnota $\mu(\theta)$	Rozptylová funkcia $V(\mu)$
Bi(n, π)	0, 1, 2 ...	1	logit $g(\mu) = \ln \frac{\mu}{1-\mu}$	$\frac{n \cdot e^\theta}{1 + e^\theta} = n \cdot \pi$	$n \cdot \pi \cdot (1 - \pi)$
Po(μ)	0, 1, 2 ...	1	logaritmus $g(\mu) = \ln(\mu)$	e^θ	μ
NBi(k, μ)	0, 1, 2 ...	1	logaritmus $g(\mu) = \ln(\mu)$	$\frac{e^\theta}{1 - k \cdot e^\theta}$	$\mu \cdot (1 + k\mu)$
N(μ, σ^2)	$(-\infty, \infty)$	σ^2	identita $g(\mu) = \mu$	$\mu = \theta$	1
$\Gamma(\mu, \alpha)$	$(0, \infty)$	$\frac{1}{\alpha}$	mocninná $g(\mu) = \frac{\mu^p}{p}$, $p = -1$	$-\frac{1}{\theta}$	μ^2
IG(μ, σ^2)	$(0, \infty)$	σ^2	mocninná $g(\mu) = \frac{\mu^p}{p}$, $p = -2$	$\frac{1}{\sqrt{-2\theta}}$	μ^3

Tab. 1 Niektoré rozdelenia z triedy exponenciálnych rozdelení

³ $a''(\theta) = \frac{d a'(\theta)}{d \theta} = \frac{d \mu}{d \theta} = V(\mu)$. Vidíme teda, že rozptyl $\text{Var}(y)$ je funkciou strednej hodnoty.

$V(\mu)$ sa nazýva *rozptylová funkcia*.

⁴ V odbornej literatúre sa možno stretnúť i s pojmom *score* pre lineárny prediktor

4. OCEŇOVANIE V NEŽIVOTNOM POISTENÍ POMOCOU GLM

Označme S_i celkové poistné plnenie i -tého poistenca. Vo všeobecnosti túto hodnotu môžeme vyjadriť ako

$$S_i = \underbrace{\sum_{k=1}^{N_i} C_{ik}}_{\text{bežné škody}} + \underbrace{p_i \cdot L_i}_{\text{veľké škody}} \quad (6)$$

kde

- N_i je počet bežných škôd (*attritional claims*) i -tého poistenca,
- C_{ik} je výška k -tej bežnej škody i -tého poistenca,
- p_i je pravdepodobnosť spôsobenia veľkej škody (*large claims*) i -tým poistencom,
- L_i je výška veľkej i -tého poistenca (ak nastane veľká škoda).

My sa v našich úvahách zameriame iba na modelovanie bežných škôd.

Nech teda pre celkové poistné plnenie poistenca platí:

$$S = \sum_{k=1}^N C_k, \quad (7)$$

kde N je celkový počet bežných škôd a C_k je výška k -tej bežnej škody.

Vzťah (7) sa nazývame aj *kolektívnym rizikom*.

Očakávané poistné plnenie potom môžeme vyjadriť v tvare:

$$E(S) = E(N) \cdot E(C) \quad (8)$$

kde $E(N)$ je škodová frekvencia a

$E(C)$ priemerné poistné plnenie.

4.1 Modelovanie škodovej frekvencie

Predpokladajme, že všetky vysvetľujúce premenné, ktoré charakterizujú poistenca tzv. *ratingové premenné*, sú kategoriálne a nadobúdajú aspoň 2 hodnoty. Ratingová premenná s k hodnotami nám rozdelí portfólio do k tarifných tried. Tieto tarifné triedy dokážeme zašifrovať pomocou $k-1$ binárnych premenných, kde všetky hodnoty budú rovné 0 pre tzv. *referenčnú triedu*. (Obvykle je referenčná trieda určená hodnotou ratingovej premennej s najväčšou expozíciou rizika.)

Lineárny prediktor pre každú tarifnú triedu je daný predpisom:

$$\boldsymbol{\eta} = \boldsymbol{\beta} \cdot \mathbf{X} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \beta_p \cdot x_p, \quad (9)$$

kde $x_1, x_2, \dots, x_p$ je postupnosť núl a jednotiek.

Nech N celkový počet bežných škôd má Poissonovo rozdelenie. Podľa tabuľky Tab.1 je linkovacou funkciou pre Poissonovo rozdelenie logaritmická funkcia.

Očakávaná (ročná) škodová frekvencia poistenca je potom rovná:

$$\exp(\boldsymbol{\eta}) = \exp(\boldsymbol{\beta} \cdot \mathbf{X}) = \exp(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots \beta_p \cdot x_p). \quad (10)$$

Teda očakávaná škodová frekvencia pre poistenca z referenčnej triedy je rovná $\exp(\beta_0)$.

Ak je $\beta_j > 0$, potom je škodová frekvencia poistenca s touto charakteristikou $\exp(\beta_j)$ -krát vyššia ako poistenca z referenčnej triedy (*ceteris paribus*).

Naopak, pokiaľ $\beta_j < 0$, potom je škodová frekvencia poistenca s touto charakteristikou $\exp(\beta_j)$ -krát nižšia ako poistenca z referenčnej triedy (*ceteris paribus*).

Číslo $\exp(\beta_j)$ modeluje efekt j -tej hodnoty ratingovej premennej.

Dostali sme teda multiplikatívny model pre škodovú frekvenciu.

Ak je poistenec charakterizovaný vektorom $\mathbf{X}_i$ pozostávajúceho z núl a jednotiek, potom očakávanú škodovú frekvenciu poistenca možno vyjadriť vzt'ahom:

$$\begin{aligned} \exp(\boldsymbol{\beta} \cdot \mathbf{X}_i) &= \exp(\beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \dots + \beta_p \cdot x_{ip}) = \\ &= \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdot \exp(\beta_2 x_{i2}) \cdot \dots \cdot \exp(\beta_p x_{ip}) = \exp(\beta_0) \cdot \prod_{j|x_{ij}=1} \underbrace{\exp(\beta_j)}_{\kappa_j} = \\ &= \exp(\beta_0) \cdot \kappa_1 \cdot \kappa_2 \cdot \dots \cdot \kappa_p, \end{aligned} \quad (11)$$

kde ak $\beta_j > 0$ (a $x_{ij} = 1$), potom $\kappa_j > 1$,

ak $\beta_j < 0$ (a $x_{ij} = 1$), potom $0 < \kappa_j < 1$.

Nech n označuje počet poistení v našom portfóliu. Nech N_i je počet poistných udalostí spôsobených i -tym poistencom $i = 1, \dots, n$. Nech d_i je riziková expozícia poistenca i (merané v rokoch). A napokon nech vektor $\mathbf{x}_i$ udáva pozorovateľné charakteristiky týkajúce sa i -teho poistenca.

Potom za predpokladu Poissonova rozdelenia počtu škôd dostávame

$$N_i \sim Po(d_i \cdot \exp(\boldsymbol{\beta} \cdot \mathbf{x}_i)), \quad i = 1, \dots, n. \quad (12)$$

tj.

$$P(N_i = k) = \exp(-d_i \cdot \exp(\boldsymbol{\beta} \cdot \mathbf{x}_i)) \cdot \frac{(d_i \cdot \exp(\boldsymbol{\beta} \cdot \mathbf{x}_i))^k}{k!}, \quad k = 0, 1, \dots$$

Očakávaný počet poistných udalostí i -teho poistenca je

$$\hat{\lambda}_i = d_i \cdot \exp(\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j \cdot x_{ij}) \quad (13)$$

Odhad parametrov modelu

Odhad parametrov modelu uskutočníme pomocou metód maximálnej vierohodnosti. Nech n_i je počet poistných udalostí i -teho poistenca (počas 1 roka). Za predpokladu nezávislosti náhodných premenných n_i a n_j , je funkcia vierohodnosti (*likelihood function*) rovná

$$\mathbf{L}(\boldsymbol{\beta}) = \prod_{i=1}^n P(N_i = n_i) = \prod_{i=1}^n \exp(-\lambda_i) \frac{\lambda_i^{n_i}}{n_i!}, \quad (14)$$

kde $\lambda_i = d_i \cdot \exp(\boldsymbol{\beta} \cdot \mathbf{x}_i)$.

Logaritmus funkcie vierohodnosti (log-likelihood function) je daný vzt'ahom

$$\begin{aligned} \ln(\mathbf{L}(\boldsymbol{\beta})) &= \sum_{i=1}^n [-\lambda_i + n_i \cdot \ln(\lambda_i) - \ln(n_i!)] = \\ &= \sum_{i=1}^n [-d_i \cdot \exp(\boldsymbol{\beta} \cdot \mathbf{x}_i)] + n_i(\boldsymbol{\beta} \cdot \mathbf{x}_i + \ln d_i) - \ln(n_i!). \end{aligned} \quad (15)$$

Označme $\hat{\beta}$ maximálne virohodný odhad parametrov.

Predpokladajme, že pre poistenie PZP MV (Poistenie Zodpovednosti za Prevádzku Motorového Vozidla) (ang. MTPL - Motor Third Party Liability) máme dáta poistenca ako riziková expozícia poistenca (exposure-to-risk), počet škôd, celkové poistné plnenie, a charakteristiku poistenca danú ratingovými premennými uvedenými v tabuľke Tab. 2.

duree	Dĺžka trvania poistenia počas jedného roka (riziková expozícia poistenca)
nbrtotc	Počet škôd
Chargtot	Celkové poistné plnenie
agecar	Vek auta: 0-1, 2-5, 6-10 , >10
fuelc	Druh pohonnej hmoty: Benzín , Nafta
sexp	Pohlavie poistenca: Muž , Žena
split	Frekvencia platenia poistného: Mesačne, Kvartálne, Polročne, Ročne
usec	Využitie auta: Súkromné , Pracovné-služobné
fleetc	Či auto patrí do flotily: Áno, Nie
sportc	Či sa jedná o športové auto: Áno, Nie
coverp	Typ poistenia: MTPL , MTPL+, MTPL+++
powerc	Výkon motoru: < 66 kW , 66-110 kW, >110 kW

Tab. 2. Premenné modelu a ich hodnoty (**tučne** vyznačené sú hodnoty referenčnej triedy)

Sformulujem si teraz modelovanú funkciu, ktorú môžeme modelovať napr. v jazyku R.

$$\begin{aligned}
 & \underbrace{glm \left(\underbrace{Data[["nbrtotc"]]}_{\text{počet pozorovaných škôd}} \sim \underbrace{offset \left(\log(Data[["duree"]]) \right)}_{\substack{\text{logaritmus} \\ \text{rizikovej expozície}}} \right)}_{\text{Vysvetľujúce premenné}} + \\
 & \underbrace{Data[["sexp"]] + Data[["fleetc"]] + Data[["sportc"]]}_{\text{Rozdelenie a linkovacia funkcia}} + \dots, \quad (16) \\
 & \text{family} = \text{poisson}(\text{link} = \text{log})).
 \end{aligned}$$

Offset sa používa v poisťovníctve ako korekcia veľkosti skupiny alebo rôznych časových období pozorovaní. V našom prípade sa offset využíva na to, že nie všetky poistky sú v platnosti po celý rok. Niektoré nadobudnú platnosť v priebehu roka, iné sú ukončené pred koncom roka. Táto skutočnosť je vyjadrená veličinou *duree*.

Predpokladajme, že $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ je maximálne virohodný odhad parametrov modelu (16).

Očakávaný počet poistných udalostí *i*-teho poistenca je: $\hat{\lambda}_i = d_i \cdot \exp(\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j \cdot x_{ij})$.

Keď vezmeme do úvahy všetky hodnoty všetkých premenných z tabuľky Tab.2., tak dostávame 4 608 tarifných tried. Položme si otázku, či sú všetky ratingové premenné štatisticky významné pre náš model. Teda poďme otestovať význam jednej alebo viacerých premenných. Za týmto účelom si zavedme niektoré pomocné pojmy.

Nech $L(\hat{\lambda})$ je virohodnosť modelu (*model likelihood*), t. j.:

$$L(\hat{\lambda}) = \prod_{i=1}^n P(N_i = \hat{\lambda}_i) = \prod_{i=1}^n \exp(-\hat{\lambda}_i) \frac{\hat{\lambda}_i^{n_i}}{n_i!} \quad (17)$$

Nech $L(\mathbf{n})$ je maximálna možná vierohodnosť (za predpokladu Poissonova rozdelenia), t. j.

$$L(\mathbf{n}) = \prod_{i=1}^n P(N_i = n_i) = \prod_{i=1}^n \exp(-n_i) \frac{n_i^{n_i}}{n_i!} \quad (18)$$

Deviance je definovaný ako:

$$D(\mathbf{n}, \hat{\lambda}) = 2 \cdot (\ln L(\mathbf{n}) - \ln L(\hat{\lambda})). \quad (19)$$

Dobrá zhoda (*goodness of fit*) modelu sa meria práve veličinou deviance – čím menšia tým lepšie.

V prípade Poissonova rozdelenia dostávame:

$$D(\mathbf{n}, \hat{\lambda}) = 2 \cdot \sum_{i=1}^n \left(n_i \cdot \ln \frac{n_i}{\hat{\lambda}_i} - (n_i - \hat{\lambda}_i) \right), \quad (20)$$

kde pre $n_i = 0$ položíme $n_i \cdot \ln \frac{n_i}{\hat{\lambda}_i} = 0$.

Náhodná premenná D má *chi*-kvadrát rozdelenie s $m - \dim(\beta)$ stupňami voľnosti, kde m je počet tarifných tried portfólia zmlúv a $\dim(\beta)$ znamená počet prvkov vektora β .

Model je považovaný za nevhodný, ak D je „príliš veľké“, tj.

$$D > \chi_{m-\dim(\beta); 1-\alpha}^2 \quad (21)$$

kde α je hladina významnosti.

Chceli by sme otestovať význam jednej alebo viacerých premenných. Teda otestujeme hypotézu

$$\begin{aligned} H_0 : \beta &= \beta_0 = (\beta_1, \beta_2, \dots, \beta_q) \text{ oproti} \\ H_1 : \beta &= \beta_1 = (\beta_1, \beta_2, \dots, \beta_q, \beta_{q+1}, \dots, \beta_p). \end{aligned} \quad (22)$$

Hypotéza H_0 nám vlastne hovorí, že $\beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0$.

Testovacou štatistikou bude rozdiel deviancí pre hypotézy H_0 a H_1 :

$$\Delta = D_0 - D_1 \sim \chi_{p-q}^2. \quad (23)$$

Hypotézu H_0 na hladine významnosti α zamietame, ak

$$\Delta > \chi_{p-q; 1-\alpha}^2. \quad (24)$$

```
> Anova(GLM_Analysis, test.statistic="Wald", type=3, singular.ok=TRUE)
Analysis of Deviance Table (Type III tests)
```

```
Response: Data[["nbrtotc"]]
```

	Df	Chisq	Pr(>Chisq)
(Intercept)	1	17239.0069	< 2.2e-16 ***
Data[["sexp"]]	1	41.1734	1.393e-10 ***
Data[["usec"]]	1	0.0347	0.852203
Data[["fleetc"]]	1	8.7443	0.003106 **
Data[["sportc"]]	1	6.5602	0.010428 *
Data[["coverp"]]	2	97.1006	< 2.2e-16 ***
Data[["fuelc"]]	1	186.5377	< 2.2e-16 ***
Data[["agecar"]]	3	59.9109	6.142e-13 ***
Data[["powerc"]]	2	36.3137	1.302e-08 ***
Data[["split"]]	3	708.0586	< 2.2e-16 ***

Vysvetľujúce
ratingové
premenné

p-hodnota

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obr. 1. Výstup ANOVA v jazyku R

Výstup ANOVA na obrázku Obr.1. nám ukazuje, že premenná „usec“ (p -hodnota je vyššia než 0,05) nie je významná pre naše modelovanie. Preto premennú „usec“ môžeme odstrániť z nášho modelu.

```
> Anova(GLM_Analysis, test.statistic="Wald", type=3, singular.ok=TRUE)
Analysis of Deviance Table (Type III tests)

Response: Data[["nbrtotc"]]
          Df      Chisq Pr(>Chisq)
(Intercept)  1 17275.4511 < 2.2e-16 ***
Data[["sexc"]]  1  41.1499  1.410e-10 ***
Data[["fleetc"]]  1   8.7150  0.003156 **
Data[["sportc"]]  1   6.5481  0.010499 *
Data[["coverp"]]  2   97.1281 < 2.2e-16 ***
Data[["fuelc"]]  1  187.7631 < 2.2e-16 ***
Data[["agecar"]]  3   59.8768  6.246e-13 ***
Data[["powerc"]]  2   36.8954  9.733e-09 ***
Data[["split"]]  3   708.1956 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obr. 2. Výstup ANOVA po odstránení premennej „usec“ v jazyku R

Na obrázku Obr.2. môžeme vidieť, podľa p -hodnoty, že už všetky ostatné premenné sú signifikantné v našom modeli.

```
> summary(GLM_Analysis)

Call:
glm(formula = Data[["nbrtotc"]] ~ offset(log(Data[["duree"]])) +
    Data[["sexc"]] + Data[["usec"]] + Data[["fleetc"]] + Data[["sportc"]] +
    Data[["coverp"]] + Data[["fuelc"]] + Data[["agecar"]] + Data[["powerc"]] +
    Data[["split"]], family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8266 -0.5306 -0.4838 -0.4242  5.3424

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.193474   0.016706 -131.297 < 2e-16 ***
Data[["sexc"]]B/Female  0.102359   0.015952   6.417 1.39e-10 ***
Data[["usec"]]B/Professional  0.006154   0.033033   0.186 0.85220
Data[["fleetc"]]B/Yes -0.128029   0.043296  -2.957 0.00311 **
Data[["sportc"]]B/Yes  0.180669   0.070538   2.561 0.01043 *
Data[["coverp"]]B/MTPL+ -0.152287   0.017150  -8.880 < 2e-16 ***
Data[["coverp"]]B/MTPL+++ -0.166963   0.024547  -6.802 1.03e-11 ***
Data[["fuelc"]]B/Gasoil  0.207886   0.015221  13.658 < 2e-16 ***
Data[["agecar"]]B/0-1  0.196461   0.033567   5.853 4.83e-09 ***
Data[["agecar"]]B/2-5 -0.054057   0.018054  -2.994 0.00275 **
Data[["agecar"]]C/>10  0.005618   0.018674   0.301 0.76352
Data[["powerc"]]B/66-110  0.092030   0.016832   5.468 4.56e-08 ***
Data[["powerc"]]C/>110  0.224668   0.070712   3.177 0.00149 **
Data[["split"]]B/Thrice  0.520837   0.023935  21.761 < 2e-16 ***
Data[["split"]]B/ Twice  0.227245   0.016709  13.600 < 2e-16 ***
Data[["split"]]C/Monthly  0.413794   0.021266  19.458 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obr. 3. Parametre modelu ŠF (výstup v jazyku R)

Obrázok Obr.3 nám ukazuje výsledné odhady parametrov (ešte pre odstránením premennej „usec“) pre všetky hodnoty našich ratingových premenných. P -hodnota nám testuje nulitu

zodpovedajúceho regresného koeficientu, nie relevantnosť ratingovej premennej, avšak v prípade, ak premenná nadobúda iba dve hodnoty (ako je to i v prípade premennej „usec“) tak je to ekvivalentné.

Ak už máme všetky ratingové premenné štatisticky významné. Položme si otázku, či je nutné pre jednotlivé ratingové premenné mať práve toľko hodnôt ako zozbiera poisťovňa pri podpise zmluvy. Napríklad, či pri výkone auta namiesto troch hodnôt <66 kW, 66-110 kW, >110 kW, by nám nestačili iba dve hodnoty: <66 kW, >66 kW.

Nech x_i (s príslušným koeficientom β_i v rovnici (10)) je binárna premenná, ktorá určuje či auto má výkon 66-110 kW, podobne nech x_j (s príslušným koeficientom β_j v rovnici (10)) je binárna premenná, ktorá určuje či auto má výkon >110 kW. Chceme overiť, či možno tieto dve skupiny áut zlúčiť do jednej.

Teda otestujeme hypotézu

$$\begin{aligned} H_0 : \beta_i &= \beta_j \quad (\text{alebo } \beta_i - \beta_j = 0) && \text{oproti} \\ H_1 : \beta_i &\neq \beta_j && (\beta_i - \beta_j \neq 0). \end{aligned} \quad (25)$$

Ukážme si to na príklade premennej vek auta „agecar“. Keďže premenná „agecar“ má štyri hodnoty: „A/6-10“, „B/0-1“, „C/2-5“ a „D/>10“, pre testovanie dostávame šesť kombinácií:

```
fit.contrast(GLM_Analysis,Data[["agecar"]],c(-1,1,0,0))
fit.contrast(GLM_Analysis,Data[["agecar"]],c(-1,0,1,0))
fit.contrast(GLM_Analysis,Data[["agecar"]],c(-1,0,0,1))
fit.contrast(GLM_Analysis,Data[["agecar"]],c(0,-1,1,0))
fit.contrast(GLM_Analysis,Data[["agecar"]],c(0,-1,0,1))
fit.contrast(GLM_Analysis,Data[["agecar"]],c(0,0,-1,1))
```

Napr. $c(-1,0,0,1)$ testuje, či možno zlúčiť dve vekové skupiny áut „A/6-10“ a „D/>10“. Ak áno, potom po zlúčení by sme zaviedli v referenčnej triede pre ratingovú premennú vek auta „agecar“ iba jednu hodnotu „A/>6“.

```
> fit.contrast(GLM_Analysis,Data[["agecar"]],c(-1,0,0,1))
                                Estimate Std. Error  z value Pr(>|z|)
Data[["agecar"]] c=(-1 0 0 1) 0.005612549 0.01867423 0.3005506 0.7637572
```

P-hodnota nám ukazuje, že by sme mali zlúčiť tieto dve skupiny áut.

Po všetkých možných zlúčeníach dostávame „iba“ 768 tarifných tried. Parametre modelu ako sú znázornené na obrázku Obr. 4.

```
Coefficients:
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                    -2.18917    0.01475 -148.437 < 2e-16 ***
Data[["sexp"]]B/Female           0.10189    0.01595   6.390 1.66e-10 ***
Data[["fleetc"]]B/Yes            -0.12821    0.04309  -2.976 0.00292 **
Data[["sportc"]]B/Yes            0.21621    0.06771   3.193 0.00141 **
Data[["coverp"]]B/More          -0.15676    0.01560 -10.049 < 2e-16 ***
Data[["fuelc"]]B/Gasoil          0.20631    0.01513  13.637 < 2e-16 ***
Data[["agecar"]]B/0-1            0.18940    0.03214   5.893 3.79e-09 ***
Data[["agecar"]]B/2-5           -0.05822    0.01677  -3.470 0.00052 ***
Data[["powerc"]]B/>=66           0.09483    0.01655   5.730 1.00e-08 ***
Data[["split"]]B/Thrice          0.52045    0.02388  21.792 < 2e-16 ***
Data[["split"]]B/ Twice          0.22683    0.01668  13.602 < 2e-16 ***
Data[["split"]]C/Monthly         0.41265    0.02116  19.497 < 2e-16 ***
```

Obr. 4. Parametre výsledného modelu pri modelovaní škodovej frekvencie (výstup v jazyku R)

Očakávaná (ročná) škodová frekvencia poistenca v referenčnej tarifnej triede je potom rovná:

$$e^{-2,18917} = 0,1120 \text{ alebo } 11,220 \, \%.$$

Túto škodovú frekvenciu môžeme považovať za základnú hodnotu, od ktorej možno odvodiť škodovú frekvenciu všetkých tarifných tried.

Ak budeme uvažovať ženu s rovnakými ostatnými parametrami, ako má muž v referenčnej tarifnej triede, tak jej očakávaná škodová frekvencia je 1,107-násobkom ($e^{0,10189} = 1,107$) základnej škodovej frekvencie, tj. je rovné $12,40\% = 1,107 \cdot 11,20\%$.

4.2 Modelovanie priemerného poistného plnenia

Ak vezmeme do úvahy štandardné-bežné poistné plnenia, na ich modelovanie sa najčastejšie využívajú rozdelenia ako *Gama*, Inverzné Gaussovo alebo Lognormalné rozdelenie. My pre modelovanie bežných škôd použijeme Gama rozdelenie.

Nech C_k je výška k -teho bežného poistného plnenia. Predpokladáme, že $C_1, C_2, \dots$ sú nezávislé. Nech každé C_k vyhovuje Gama rozdeleniu $C_k \sim \Gamma(\mu_k, \alpha)$

$$\mu_k = E(C_k) = \exp(\boldsymbol{\beta} \cdot \mathbf{X}_k), \quad (26)$$

$$\sigma_k^2 = Var(C_k) = \frac{(\exp(\boldsymbol{\beta} \cdot \mathbf{X}_k))^2}{\alpha}. \quad (27)$$

Nech n_i je počet škôd i -tého poistenca a nech $c_{i1}, c_{i2}, \dots, c_{i,n_i}$ sú príslušné poistné plnenia plynúce z týchto škôd.

Funkcia vierohodnosti spojená so všetkými plneniami má potom tvar

$$L(\boldsymbol{\beta}) = \prod_{i|n_i>0} \prod_{k=1}^{n_i} f(c_{ik}|x_i, \alpha) = \prod_{i|n_i>0} \prod_{k=1}^{n_i} \left(\frac{1}{\Gamma(\alpha)} \cdot \left(\frac{\alpha \cdot c_{ik}}{\mu_i} \right)^\alpha \cdot \exp\left(-\frac{\alpha \cdot c_{ik}}{\mu_i}\right) \right) \cdot \frac{1}{c_{ik}}. \quad (28)$$

Podobne ako pri škodovej frekvencii pre výpočet $\hat{\boldsymbol{\beta}}$ použijeme maximálne vierohodný odhad. Potom

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = 0 \quad \Leftrightarrow \quad \sum_{i|n_i>0} \sum_{k=1}^{n_i} x_{ij} \cdot \left(1 - \frac{c_{ik}}{\mu_i} \right) = 0 \quad (29)$$

Sformulujem si opäť modelovanú funkciu, ktorú budeme modelovať napr. v jazyku R.

$$\begin{aligned} & \underbrace{glm(Data[["MeanCharge"]])}_{\text{priemerné poist.plnenie}} \sim \\ & \underbrace{Data[["sexc"]] + Data[["fleetc"]] + Data[["sportc"]] + \dots}_{\text{Vysvetľujúce premenné}}, \\ & \underbrace{family = Gamma(link = log)}_{\text{Rozdelenie a linkovacia funkcia}}, \underbrace{weights = Data[["nbrtotc"]]}_{\text{váha = počet škôd}} \end{aligned} \quad (30)$$

Ďalší postup je podobný ako pri modelovaní škodovej frekvencie.

Po získaní parametrov $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ opäť určíme, ktoré premenné sú štatisticky významné pre výšku poistného plnenia.

Postupujeme rovnako ako pri modelovaní škodovej frekvencie. Určíme, nielen to, ktoré ratingové premenné sú významné ale i to, ktoré hodnoty jednotlivých ratingových premenných možno zlúčiť pod jednu hodnotu.

Na úvod eliminujeme extrémne vysoké škody $L = \text{DataCosts}[["chargetot"]] < 100000$ a eliminujeme taktiež dáta s nulovými škodami $L = \text{DataCosts}[["chargetot"]] != 0$.

Po testovaní relevantnosti všetkých premenných a po možných zlúčeníach ich hodnôt dostávame výsledné parametre modelu, ako nám ukazuje výstup v jazyku R, Obr. 5.

Priemerné poistné plnenie pri škode spôsobenej poistencom v referenčnej tarifnej triede je potom rovná:

$$e^{7,20832} = 1\,350,62.$$

Toto poistné plnenie považujeme opäť za základnú hodnotu, od ktorej možno odvodiť poistné plnenie škôd všetkých tarifných tried.

Ak budeme uvažovať poistenca s novým autom (vek 0-1 rok), ktorý má všetky ostatné ukazovatele zhodné s poistencom v referenčnej tarifnej triede, tak jeho priemerné poistné plnenie je 1,27-násobkom ($e^{0,23889} = 1,27$) základného poistného plnenia, tj. je rovné $1\,350,62 \cdot 1,27$.

```
> summary(GLM_AnalysisCosts)

Call:
glm(formula = DataCosts[["MeanCharge"]] ~ DataCosts[["coverp"]] +
    DataCosts[["agecar"]], family = Gamma(link = log), weights = DataCosts[["nbrtotc"]])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4510  -1.6530  -0.8469   0.0551  11.8843

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.20832    0.02547  282.965  < 2e-16 ***
DataCosts[["coverp"]]B/MTPL+ -0.24236    0.04616  -5.251 1.53e-07 ***
DataCosts[["coverp"]]B/MTPL+++ 0.12142    0.06244   1.945  0.05183 .
DataCosts[["agecar"]]B/0-1    0.23889    0.09026   2.647  0.00813 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obr. 5. Parametre modelu pri modelovaní výšky poistného plnenia (výstup v jazyku R)

5. ZÁVER

Ak si zhrnieme výsledky modelovania škodovej frekvencie a modelovania očakávanej výšky škody, tak pre očakávané poistné plnenie, čo pre nás reprezentuje netto poistné, dostávame:

$$\begin{aligned} E(S) &= E(N) \cdot E(C) = \exp(\boldsymbol{\beta} \cdot \mathbf{X}_i) \cdot \exp(\boldsymbol{\gamma} \cdot \mathbf{X}_i) = \exp[(\boldsymbol{\beta} + \boldsymbol{\gamma}) \cdot \mathbf{X}_i] = \\ &= \exp(\beta_0 + \gamma_0) \cdot \prod_{j|x_{ij}=1} \underbrace{\exp(\beta_j + \gamma_j)}_{t_j} = \exp(\beta_0 + \gamma_0) \cdot t_1 \cdot t_2 \cdot \dots \cdot t_p, \end{aligned} \quad (34)$$

kde $P_0 = \exp(\beta_0 + \gamma_0)$ je základné netto poistné (netto poistné referenčnej triedy), od ktorej sú odvodené ostatné hodnoty buď prirážkou, ak $t_j > 1$ alebo zľavou, ak $0 < t_1 < 1$.

Variable	Modality	Frequency		Costs		Pure premium for standard claims	
		Estimate	Relativities	Estimate	Relativities	Estimate	Relativities
Intercept		-2,18917	11,2%	7,20832	1 350,62	5,01915	151,28
Sexp	Female	0,10189	110,7%	0	100,0%	0,10189	110,7%
Fleetc	Yes	-0,12821	88,0%	0	100,0%	-0,12821	88,0%
Sportc	Yes	0,21621	124,1%	0	100,0%	0,21621	124,1%
Coverp2	MTPL+	-0,15676	85,5%	-0,24236	78,5%	-0,39912	67,1%
	MTPL+++	-0,15676	85,5%	0,12142	112,9%	-0,03534	96,5%
Fuelc	Gasoil	0,20631	122,9%	0	100,0%	0,20631	122,9%
Agecar2	0-1	0,1894	120,9%	0,23889	127,0%	0,42829	153,5%
	2-5	-0,05822	94,3%	0	100,0%	-0,05822	94,3%
Powerc	>=66	0,09483	109,9%	0	100,0%	0,09483	109,9%
Split	Monthly	0,41265	151,1%	0	100,0%	0,41265	151,1%
	Thrice	0,52045	168,3%	0	100,0%	0,52045	168,3%
	Twice	0,22683	125,5%	0	100,0%	0,22683	125,5%

Tab. 3. Výsledky modelu GLM

Počiatočný model so všetkými ratingovými premennými a ich hodnotami nám vytvárajú 4 608 tarifných skupín ($4 \cdot 2 \cdot 2 \cdot 4 \cdot 2 \cdot 2 \cdot 3 \cdot 3$). Náš model je schopný určiť poistné pre každú tarifnú triedu, i takú v ktorej má poistiteľ málo štatistických dát, ba možno nemá žiadne skúsenosti v doterajšej histórii pôsobenia na trhu.

Základné netto poistné vo výške **151,28** p.j. nám zvyšujú prirážky – vyznačené červenou farbou a znižujú zľavy – vyznačené zelenou farbou.

Tak napríklad poistenec, ktorý sa od jedinca v referenčnej tarifnej triede líši iba tým, že má športové auto vo veku 2-5 rokov, bude mať netto poistné vo výške:

$$151,28 \cdot 1,241 \cdot 0,943 = 177,18 \text{ p.j.}$$

LITERATÚRA

- [1] DENUIT, M., MARÉCHAL, X., PITREBOIS, S., & WALHIN, J.-F., *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Wiley, New York, 384 s., 2007, ISBN 978-0-470-02677-9.
- [2] McCULLAGH, P., NELDER, J.A., *Generalized Linear Models* (2nd edition), London, Chapman & Hall, 511 s., 1989, ISBN 0-412-31760-5.
- [3] PACÁKOVÁ, V., *Štatistika pre poistnú prax*, Bratislava, Statis, 187 s., 1999, ISBN 80-85659-14-X.
- [4] R Core Team. The R project for statistical computing. URL <https://www.r-project.org>, 2021
- [5] ŠOLTÉS, E., *Regresná korelačná analýza s aplikáciami*. Vyd. 1. Bratislava: Iura Edition, 287 s., 2008, ISBN 978-80-8078-163-7.

KONTAKTNÉ ÚDAJE

Gogola, Ján, RNDr., PhD., Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemska cesta 1, 852 35 Bratislava, e-mail: jan.gogola@euba.sk

Zelinová, Silvia, Ing., PhD., Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemska cesta 1, 852 35 Bratislava, e-mail: silvia.zelinova@euba.sk